Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

Zheyang Shen

Spectral kernels for Gaussian processes

Master's Thesis
Espoo, February 18, 2019

Supervisor: Professor Samuel Kaski, Aalto University
Advisor: Markus Heinonen, Ph.D

Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

ABSTRACT OF MASTER'S THESIS

Author:	Zheyang Shen		
Title:	Spectral kernels for Gaussian processes		
Date:	February 18, 2019	**Pages:**	vii + 42
Major:	Machine Learning, Data Science and Artificial Intelligence	**Code:**	SCI3044
Supervisor:	Professor Samuel Kaski		
Advisor:	Markus Heinonen, Ph.D		

Gaussian processes are flexible distributions over functions, which provide a non-parametric nonlinear Bayesian regression framework. The *covariance kernel*, an operator determining the similarity between two points, is central to every Gaussian process model, for it encodes prior knowledge of the function being inferred.

In this work, we extend the expressive power of Gaussian process models by proposing two families of kernels, over which an efficient search of expressive hidden representations of data is possible. The *harmonizable mixture kernel* (HMK) is a theoretically sound approach, to derive a parametric kernel by taking the Fourier transform of a generalized spectral density, modeled by a Gaussian mixture model. The *convolutional spectral kernel* (CSK) is a nonparametric kernel generalizing HMK, derived from taking the convolution of two spectral mixture kernel feature maps. We show that the two classes of kernels theoretically exhibit high levels of expressiveness, and we introduce Wigner distribution functions as a useful tool to interpret kernels.

We also study efficient inference specially designed for the two new kernel families. We propose *variational Fourier features* (VFF), an inter-domain sparse inference approach utilizing the generalized spectral density.

Experiments are extensively conducted for the two kernels and one new inference methods. We demonstrate experimentally that HMK interpolates between local patterns, and VFF offers a robust framework for learning kernel hyperparameters. We show that CSK can extract complex patterns using a nonparametric approach, with the added advantage of adapting spectral frequencies for each pair of data points.

Keywords:	Gaussian processes, kernel methods, non-stationary covariances
Language:	English

Declaration

I, Zheyang Shen, being a candidate for the Master of Science in Machine Learning, Data Science and Artificial Intelligence, hereby declare that this thesis and the work described in it are my own work, unaided except as may be specified below, and that the report does not contain material that has already been used to any substantial extent for a comparable purpose.

The content of this thesis is being partially published in the Proceedings of Artificial Intelligence and Statistics, 2019, with title *harmonizable mixture kernels with variational Fourier features*.

Zheyang Shen
February 18, 2019

Acknowledgements

While this thesis bears my name as the author, it is indeed a joint work between me and my advisors. I would like to thank my supervisor Professor Samuel Kaski and my advisor Dr. Markus Heinonen for their continued support and guidance of my research and supervision over the writing of my thesis. It is safe to say the body of my work would have been utterly futile were it not for the hard work of my advisors. I am also very fortunate to have worked with the Probabilistic Machine Learning Group of Aalto University, which provides invaluable guidance towards the completion of my master thesis. I also would like to thank Aalto CS-IT and Science-IT for providing computational resources necessary for my experiments.

On a personal note, I would like to thank my friends, family and colleagues for their assistance and love during the process. I appreciate the open-mindedness and moral support from my parents, Pingping Feng and Hongchang Shen, and I am grateful that my friends Yue, Cuong, Guangyi, Yuanhao and Qianyun fill my life with sporadic moments of happiness. Furthermore, thank my grandparents for all their support throughout my life.

Espoo, February 18, 2019

Zheyang Shen

Abbreviations and Acronyms

GP	Gaussian process
SE	Squared exponential (Gaussian) kernel
HMK	Harmonizable mixture kernel
CSK	Convolutional spectral kernel
VFF	Variational Fourier features
RFF	Random Fourier features
SM	Spectral mixture kernel [39]
PS	k^{NS} proposed by Paciorek and Schervish [19]
GSM	Generalized spectral mixture kernel [25]
GSD	Generalized spectral density [30]

Contents

Chapter 1

Introduction

Machine learning is fundamentally about pattern recognition. A good machine learning model can, in principle, not only help human analyze data, but also learn hidden representations in data, thus automating the learning and decision making process [39].

Kernel method is one of the cornerstones of machine learning and pattern recognition. Kernels, as a measure of similarity between two objects, depart from the common linear hypotheses by allowing for complex nonlinear patterns [36]. In a Bayesian framework, kernels are interpreted probabilistically as covariance functions of random processes, such as for the Gaussian processes in Bayesian nonparametrics.

Gaussian processes (GP) have been of increasing interest in the machine learning community. As rich distributions over functions, GPs are noted for their connection to Bayesian neural networks [38], as well as their tractability, robustness to overfitting and scalability [24]. Properties of likely functions drawn from a GP, e.g., periodicity and differentiability, are determined by a positive definite *covariance kernel*. The choice of kernel is thus central to any GP models for it encodes prior knowledge of the function being inferred.

Despite the rich nonlinearity permitted by kernels, GPs are rarely used as expressive statistical tools, but instead as simple smoothing devices mainly intended for interpolation. Squared exponential (SE) kernels are used by default, which includes a set of fixed basis functions encoding only global and monotonic covariance patterns. Smoothing devices are not sufficient to compare with the automatic feature extraction performed by neural networks.

Various efforts have contributed to the automation of pattern recognition inside a GP framework, or more exactly, the construction of expressive kernels capable of extracting hidden representations in data. Early works explored the possiblity of local monotonic covariances by convolving more expressive feature maps [8, 19]. Recent works construct expressive kernels mainly by

two different approaches, namely by mimicking a manual search for best kernel forms by adding, multiplying and composing simple kernel forms [1, 4, 9, 24, 33] and by modeling the spectral representations [21, 25, 28, 35, 39]. However, previous approaches are not without their pitfalls, namely tendencies to overfit [1, 4, 9, 21, 39], lack of interpretabilities [1, 4, 9, 25, 33], stationarity constraints [21, 35, 39], ad-hoc choices of certain kernel forms [1, 4, 9, 25, 33], and inclusion of invalid kernels [28].

In this thesis, we seek to overcome the shortcomings and present a unified view over various perspectives provided by previous works by proposing a new, theoretically sound framework that bridges and generalizes them. Our main contributions include

- We introduce *harmonizability*, a term previously used only in statistics literature, into the field of machine learning. Harmonizable kernels generalize stationary kernels by allowing for non-stationarity, while remaining interpretable with their *generalized* spectral representations.

- We propose the practical *harmonizable mixture kernels* (HMK), a class of kernels dense in the set of harmonizable covariances with generalized spectral distributions.

- We propose *convolutional spectral kernels*, a spectral kernel family taking flexible functions as input. We demonstrate its expressiveness by showing its ability of including various previous kernels as special cases.

- We propose *variational Fourier features*, an inter-domain GP inference framework for GPs equipped with HMK. Functions drawn from such GP priors have a well-defined Fourier transform, a desirable property not found in stationary GPs.

Chapter 2

Two sets of spectral kernels

In this chapter, we introduce two sets of spectral kernels, with theoretical analysis of each kernel. We introduce necessary background information in 2.1, namely on kernel methods and two approaches to construct expressive kernels relevant to this thesis. We introduce *harmonizability* and present certain subclass of harmonizable kernels in 2.2. We construct *harmonizable mixture kernels*, a new kernel family that spans harmonizable covariances in 2.3. We construct *convolutional spectral kernels*, an interpretable nonparametric spectral kernel in 2.4. We explore the expressiveness of the new kernel families in 2.5. We seek interpretation of spectral kernels in 2.6. Finally, we empirically demonstrate the expressiveness of spectral kernels using kernel recovery experiments in 2.8.

2.1 Background

This section is a short tutorial about kernel methods and two methods to construct expressive kernels. For notational consistency, we denote the input domain by $\mathcal{X}$: $\mathbf{x} \in \mathcal{X}$, and we mostly consider vectorial input: $\mathcal{X} = \mathbb{R}^D$.

2.1.1 Kernel method

Kernel method studies similarity measures between objects. In this section, we demonstrate an equivalence between *positive definite kernels* with feature mappings.

Positive definite kernels $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{K}$, where $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$, stem from positive definite Gram matrices.

Definition 1. Let $\mathcal{X}$ be a nonempty set, a function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{K}$ which for all $m \in \mathbb{N}$ and $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathcal{X}$ gives rise to a positive definite $m \times m$ matrix

$\mathbf{K}$, where $\mathbf{K}_{ij} \triangleq k(\mathbf{x}_i, \mathbf{x}_j)$. Such function k is a *positive definite kernel*, or we shall simply refer to the function k as a *kernel*.

The positive definiteness of kernels implies *positivity on the diagonal* and *symmetry*:

$$k(\mathbf{x}, \mathbf{x}) \geq 0, \forall \mathbf{x} \in \mathcal{X}, \tag{2.1}$$

$$k(\mathbf{x}, \mathbf{x}') = \overline{k(\mathbf{x}', \mathbf{x})}, \tag{2.2}$$

where $\bar{g}$ defines complex conjugate when k is complex-valued .

Given a positive definite kernel k, we define a map $\boldsymbol{\phi} : \mathcal{X} \mapsto \mathbb{K}^{\mathcal{X}} \triangleq \{f : \mathcal{X} \mapsto \mathbb{K}\}$, via

$$\begin{aligned} \boldsymbol{\phi} :& \mathcal{X} \mapsto \mathbb{K}^{\mathcal{X}}, \\ & \mathbf{x} \mapsto k(\cdot, \mathbf{x}). \end{aligned} \tag{2.3}$$

It was proved[1] that the "feature map" $\boldsymbol{\phi}$ defines a dot product on the vector space containing the image of $\boldsymbol{\phi}$ such that, the dot product effectively reproduces the kernel:

$$k(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}') \rangle. \tag{2.4}$$

The mathematical theory of equivalence between kernels and dot product in a linear feature space has a profound impact on machine learning. It allows machine learning algorithms to depart from common linear hypotheses, and operate in a nonlinear projection of possibly infinite dimensional feature space using functions satisfying the positive definite property, and define measures of similarity for vastly different sets of $\mathcal{X}$.

Kernels usually come with *kernel hyperparameters*, a set of parameters that makes the kernel function form valid. *Kernel learning* refers to the learning of those hyperparameters – it is an attempt to extract hidden representations of data encoded within the projected feature map. When we are given a class of kernels with a flexible enough feature map, we can extract patterns using kernel learning.

2.1.2 Stationary kernels

Stationary kernels are an important and well-studied subset of kernels. A *stationary kernel* is a kernel whose value is a function of $\boldsymbol{\tau} = \mathbf{x} - \mathbf{x}'$, i.e., it is invariant to translation of inputs:

$$k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x} - \mathbf{x}'), \tag{2.5}$$

$$k(\mathbf{x} + \mathbf{z}, \mathbf{x}' + \mathbf{z}) = k(\mathbf{x}, \mathbf{x}'). \tag{2.6}$$

[1]The detailed proof is in section 2.2.2 in Smola and Schölkopf [31]

Stationary kernels can be fully characterized by finite measures using Bochner's theorem [2]. Bochner's theorem defines a one-to-one mapping from stationary kernels to finite measures via a Fourier transform.

Theorem 1. *(Bochner) A complex-valued function* $k : \mathbb{R}^D \times \mathbb{R}^D \mapsto \mathbb{C}$ *is a stationary kernel if and only if it can be represented as*

$$k(\boldsymbol{\tau}) = \int_{\mathbb{R}^D} e^{2i\pi \boldsymbol{\xi}^\top \boldsymbol{\tau}} \psi_k(d\boldsymbol{\xi}), \tag{2.7}$$

where ψ_k *is a positive finite measure, denoted as the* spectral distribution *of* k*.*

We can construct any stationary kernel by defining a finite measure, and then using the inverse Fourier transform. More specifically, we can construct expressive classes of kernels using mixtures of distributions for ψ_k. For example:

- Finite pure point measures give the *sparse spectrum kernel* [21]:

$$k_{\mathrm{SS}}(\boldsymbol{\tau}) = \sum_{q=1}^{Q} w_q^2 \exp(2i\pi \boldsymbol{\omega}_q^\top \boldsymbol{\tau}), \tag{2.8}$$

$$\psi_{k_{\mathrm{SS}}}(\boldsymbol{\xi}) = \sum_{q=1}^{Q} w_q^2 \delta_{\boldsymbol{\xi}=\boldsymbol{\omega}_q}. \tag{2.9}$$

- Gaussian mixtures give the *spectral mixture kernel* [39]:

$$k_{\mathrm{SM}}(\boldsymbol{\tau}) = \sum_{q=1}^{Q} w_q^2 \exp(-2\pi^2 \boldsymbol{\tau} \boldsymbol{\Sigma}_q \boldsymbol{\tau} + 2i\pi \boldsymbol{\omega}_q^\top \boldsymbol{\tau}), \tag{2.10}$$

$$\psi_{k_{\mathrm{SM}}}(\boldsymbol{\xi}) = \sum_{q=1}^{Q} w_q^2 \mathcal{N}(\boldsymbol{\xi} | \boldsymbol{\omega}_q, \boldsymbol{\Sigma}_q). \tag{2.11}$$

We will discuss the expressiveness of the two classes of kernels in 2.5. Note that k_{SS} encodes a finite-dimensional feature map, which renders it functionally equivalent with a finite basis expansion with trigonometric functions. Expressive stationary kernels are appealing in the sense that they offer a flexible feature map with the inductive bias of stationarity.

2.1.3 Convolutional kernels

Convolutional kernels [13, 20] stem from an explicit construction of feature maps. A covariance function $C(\cdot,\cdot)$ can be constructed by convolving two kernel functions centered respectively at $\mathbf{x}_i$ and $\mathbf{x}_j$:

$$C(\mathbf{x}_i,\mathbf{x}_j) = \int_{\mathbb{R}^D} K_{\mathbf{x}_i}(\mathbf{u})\overline{K}_{\mathbf{x}_j}(\mathbf{u})\,\mathrm{d}\mathbf{u}, \tag{2.12}$$

where $K_{\mathbf{x}}(\mathbf{u})$ is a kernel function centered around $\mathbf{x}$. The positive definiteness of $C(\cdot,\cdot)$ is straightforward given that the convolution is consistent with a dot product of feature maps.

Paciorek [20] derives a closed-form generalization of *squared exponential kernel* by defining $K_{\mathbf{x}_i}(\mathbf{u}) = \exp\left(-\frac{1}{2}(\mathbf{x}_i-\mathbf{u})^\top\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\mathbf{u})\right)$. The derivation is straightforward once we consder the Fourier transform of $K_{\mathbf{x}_i}(\mathbf{u})$. After proper normalization, we get a non-stationary correlation function:

$$R(\mathbf{x}_i,\mathbf{x}_j) = \frac{2^{D/2}|\boldsymbol{\Sigma}_i|^{1/4}|\boldsymbol{\Sigma}_j|^{1/4}}{|\boldsymbol{\Sigma}_i+\boldsymbol{\Sigma}_j|^{1/2}}\exp\left(-(\mathbf{x}_i-\mathbf{x}_j)^\top\left(\frac{\boldsymbol{\Sigma}_i+\boldsymbol{\Sigma}_j}{2}\right)^{-1}(\mathbf{x}_i-\mathbf{x}_j)\right). \tag{2.13}$$

It is straightforward to see that (2.13) recovers the squared exponential covariance $R_{\mathrm{SE}}(\mathbf{x}_i,\mathbf{x}_j) = \exp\left(-(\mathbf{x}_i-\mathbf{x}_j)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\mathbf{x}_j)\right)$ when $\boldsymbol{\Sigma}_i \equiv \boldsymbol{\Sigma}$. The same kernel form where $\boldsymbol{\Sigma}_i$ is also derived in Gibbs [8].

Construction of expressive kernels using convolution often gives a non-stationary kernel with a functional parameter – (2.13) gives a non-stationary correlation with input dependent covariance. However, this kernel form is limited in that it is positive-valued and monotonic.

2.2 Harmonizable covariances

In this section we introduce *harmonizability*, a generalization of stationarity previously not considered in the machine learning literature. We first define harmonizable kernel, and then analyze two existing special cases of harmonizable kernels, stationary and locally stationary kernels.

A harmonizable kernel [16, 18, 41] is a kernel with a *generalized spectral distribution* defined by a generalized Fourier transform:

Definition 2. A complex-valued bounded continuous kernel $k : \mathbb{R}^D \times \mathbb{R}^D \mapsto \mathbb{C}$ is *harmonizable* when it can be represented as

$$k(\mathbf{x},\mathbf{x}') = \int_{\mathbb{R}^D\times\mathbb{R}^D} e^{2i\pi(\boldsymbol{\omega}^\top\mathbf{x}-\boldsymbol{\xi}^\top\mathbf{x}')}\mu_{\Psi_k}(\mathrm{d}\boldsymbol{\omega},\mathrm{d}\boldsymbol{\xi}), \tag{2.14}$$

where μ_{Ψ_k} is the Lebesgue-Stieltjes measure associated to some positive definite function $\Psi_k(\boldsymbol{\omega}, \boldsymbol{\xi})$ with bounded variations.

The positive definite measure induced by function Ψ_k is defined as the generalized spectral distribution of the kernel, and when μ_{Ψ_k} is twice differentiable, the derivative $S_k(\boldsymbol{\omega}, \boldsymbol{\xi}) = \dfrac{\partial^2 \Psi_k}{\partial \boldsymbol{\omega} \partial \boldsymbol{\xi}}$ is defined as *generalized spectral density* (GSD).

Harmonizable kernel is a very general class in the sense that it contains a large portion of bounded, continuous kernels with only a handful of (somewhat pathological) exceptions [41].

2.2.1 Comparison with Bochner's theorem

One can clearly see the similarity between the definition of harmonizable kernel (2.14) and the Fourier transform specified by the Bochner's theorem (2.7): both of them draw a connection between measures and kernels using a Fourier transform.

The harmonizable definition is indeed a generalization of Bochner's theorem. When the mass of the measure μ_Ψ is concentrated on the diagonal $\boldsymbol{\omega} = \boldsymbol{\xi}$, the generalized inverse Fourier transform devolves into an inverse Fourier transform with respect to $\boldsymbol{\tau} = \mathbf{x} - \mathbf{x}'$, and therefore recovers the exact form in Bochner's theorem.

However, there are some key distinctions between harmonizable kernels and Bochner's theorem. While Bochner's theorem specifies a positive finite measure ψ_k, the *generalized spectral distribution* Ψ_k is a complex-valued measure, with subsets assigned to complex numbers.

Bochner's theorem inherently specifies a feature map, while such feature map is notably absent from the harmonizable definition. The integral (2.7) can be seen as an expectation after normalizing ψ_k as a probability measure:

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= \int_{\mathbb{R}^D} e^{2i\pi \boldsymbol{\xi}^\top (\mathbf{x} - \mathbf{x}')} \psi_k(\mathrm{d}\boldsymbol{\xi}) \\ &= \sigma_{\psi_k}^2 \mathbb{E}_{\boldsymbol{\xi} \sim p(\psi_k)} \left(\exp(2i\pi \boldsymbol{\xi}^\top \mathbf{x}) \overline{\exp(2i\pi \boldsymbol{\xi}^\top \mathbf{x}')} \right). \end{aligned} \tag{2.15}$$

The expectation representation (2.15) gives rise to the random Fourier features [23], an approximation of feature map derived from sampling the spectral distribution ψ_k. While Bochner's theorem allows easy sampling and straightforward feature map representation, neither is apparent for harmonizable kernels: the complex measure Ψ_k prevents a sampling paradigm, and the exponential term $e^{2i\pi(\boldsymbol{\xi}^\top \mathbf{x} - \boldsymbol{\omega}^\top \mathbf{x}')}$ no longer follows a dot product structure.

2.2.2 Locally stationary kernels

As a generalization of stationary kernels, the locally stationary kernels [30] are a simple yet unexplored concept in machine learning. A locally stationary kernel is a stationary kernel multiplied by a sliding power factor:

$$k_{LS}(\mathbf{x}, \mathbf{x}') = k_1\left(\frac{\mathbf{x}+\mathbf{x}'}{2}\right) k_2(\mathbf{x}-\mathbf{x}'). \tag{2.16}$$

where $k_1 : \mathbb{R}^D \mapsto \mathbb{R}_{\geq 0}$ is an arbitrary nonnegative function, and $k_2 : \mathbb{R}^D \mapsto \mathbb{C}$ is a stationary kernel. k_1 is a function of the *centroid* between $\mathbf{x}$ and $\mathbf{x}'$, describing the scale of covariance on a global structure, while k_2 as a stationary covariance describes the local structure [7]. It is straightforward to see that locally stationary kernels reduce into stationary kernels when k_1 is constant.

Integrable locally stationary kernels are of particular interest because they are harmonizable with a GSD. Consider a locally stationary Gaussian kernel (LSG) defined as a SE kernel multiplied by a Gaussian density on the centroid $\widetilde{\mathbf{x}} = (\mathbf{x}+\mathbf{x}')/2$. Its GSD can be obtained using the generalized Wiener-Khintchin relations [30].

$$k_{\mathrm{LSG}}(\mathbf{x}, \mathbf{x}') = e^{-2\pi^2 \widetilde{\mathbf{x}}^\top \boldsymbol{\Sigma}_1 \widetilde{\mathbf{x}}} e^{-2\pi^2 \boldsymbol{\tau}^\top \boldsymbol{\Sigma}_2 \boldsymbol{\tau}}, \tag{2.17}$$

$$S_{k_{\mathrm{LSG}}}(\boldsymbol{\omega}, \boldsymbol{\xi}) = \mathcal{N}\left(\frac{\boldsymbol{\omega}+\boldsymbol{\xi}}{2}\middle| 0, \boldsymbol{\Sigma}_2\right) \mathcal{N}\left(\boldsymbol{\omega}-\boldsymbol{\xi}\middle| 0, \boldsymbol{\Sigma}_1\right). \tag{2.18}$$

2.3 Harmonizable mixture kernels

In this section we propose a novel *harmonizable mixture kernel*, a family of kernels dense in harmonizable covariance functions. Our construction is inspired by the spectral mixture kernel [39].

Consider the simple one-dimensional case $\mathcal{X} = \mathbb{R}$. The integral (2.14) can thus be discretized over a grid of ωs: $\omega_0 < \omega_1 < \cdots < \omega_m$:

$$\begin{aligned}\int_{\mathbb{R}\times\mathbb{R}} e^{2i\pi(\omega x - \xi x')} \mu_{\Psi_k}(\mathrm{d}\omega, \mathrm{d}\xi) &\approx \sum_{i=0}^{m-1}\sum_{j=0}^{m-1} e^{2i\pi(\omega_i x - \omega_j x')} \Psi([\omega_i, \omega_{i+1}] \times [\omega_j, \omega_{j+1}]) \\ &= \boldsymbol{\phi}(x)^\dagger \mathbf{B} \boldsymbol{\phi}(\mathbf{x}'),\end{aligned} \tag{2.19}$$

where $\boldsymbol{\phi}(x)_i = e^{2i\pi\omega_i x}$, $\mathbf{B}_{ij} = \Psi([\omega_i, \omega_{i+1}] \times [\omega_j, \omega_{j+1}])$, and $\boldsymbol{\phi}(x)^\dagger$ denotes vector Hermitian. We can see that the Darboux sum of the integral is equivalent to a finite basis expansion $\boldsymbol{\phi}(x)$, combined with an inner product specified

by a positive definite matrix $\mathbf{B}$. The kernel form $k_{\text{GSS}} = \boldsymbol{\phi}(x)^\dagger \mathbf{B} \boldsymbol{\phi}(\mathbf{x}')$ is a generalization of the sparse spectrum kernel.

Spectral mixture kernel can be viewed as a sparse spectrum kernel multiplied by an SE kernel. This perspective is partially discussed by Samo and Roberts [28]. We can apply similar logic to constructing a harmonizable kernel with infinite-dimensional feature map. Consider a positive definite, continuous and integrable kernel $h(\cdot, \cdot)$, the following form is a valid kernel encoding a possibly infinite-dimensional feature map:

$$k(x, x') = h(\gamma x, \gamma x') \boldsymbol{\phi}(x)^\dagger \mathbf{B} \boldsymbol{\phi}(\mathbf{x}') \to \boldsymbol{\phi}(x)^\dagger \mathbf{B} \boldsymbol{\phi}(x') \text{ as } \gamma \to 0^+. \tag{2.20}$$

Using the added flexibility of the kernel h, we can construct a generalized form of the *harmonizable mixture kernel* in a multidimensional setting:

$$k_p(\mathbf{x}, \mathbf{x}') = h(\mathbf{x} \circ \boldsymbol{\gamma}_p, \mathbf{x}' \circ \boldsymbol{\gamma}_p) \boldsymbol{\phi}_p(\mathbf{x})^\dagger \mathbf{B}_p \boldsymbol{\phi}_p(\mathbf{x}'), \tag{2.21}$$

where $\circ$ denotes the Hadamard (pointwise) product of two vectors, $\phi_{pq}(\mathbf{x}) = e^{2i\pi \boldsymbol{\mu}_{pq}^\top \mathbf{x}}, q = 1, \ldots, Q_p$ are sinusoidal feature maps, $\mathbf{B}_p \succeq \mathbf{0}_{Q_p}$ are spectral amplitudes, $\boldsymbol{\gamma}_p \in \mathbb{R}_+^D$ are input scalings, and $\boldsymbol{\mu}_{pq} \in \mathbb{R}^D$ are frequencies. In a concrete setting, we propose the locally stationary Gaussian kernel k_{LSG}(2.17) as a suitable candidate for h.

The multiplication of an integrable kernel h renders the kernel k_p (2.21) local. We can add to the flexibility of (2.21) by adding shifts of the input space:

$$k_{\text{HM}}(\mathbf{x}, \mathbf{x}') = \sum_{p=1}^{P} k_p(\mathbf{x} - \mathbf{x}_p, \mathbf{x}' - \mathbf{x}_p). \tag{2.22}$$

2.3.1 Spectral representations

Utilizing the harmonizability theory, we can construct spectral representation of the proposed *harmonizable mixture kernel* (2.22) when $h = k_{\text{LSG}}$. The derivation is straightforward using the generalized Fourier transform (2.14). The generalized spectral density (GSD) of k_{HM} takes the following form:

$$S_{k_{\text{HM}}}(\boldsymbol{\omega}, \boldsymbol{\xi}) = \sum_{p=1}^{P} S_{k_p}(\boldsymbol{\omega}, \boldsymbol{\xi}) e^{-2i\pi \mathbf{x}_p^\top (\boldsymbol{\omega} - \boldsymbol{\xi})}, \tag{2.23}$$

$$S_{k_p}(\boldsymbol{\omega}, \boldsymbol{\xi}) = \frac{1}{\prod_{d=1}^{D} \gamma_{pd}^2} \sum_{1 \le i,j \le Q_p} b_{pij} S_{pij}(\boldsymbol{\omega}, \boldsymbol{\xi}), \tag{2.24}$$

$$S_{pij}(\boldsymbol{\omega}, \boldsymbol{\xi}) = S_{k_{\text{LSG}}}((\boldsymbol{\omega} - \boldsymbol{\mu}_{pi}) \oslash \boldsymbol{\gamma}_p, (\boldsymbol{\xi} - \boldsymbol{\mu}_{pj}) \oslash \boldsymbol{\gamma}_p), \tag{2.25}$$

where b_{pij} denotes the ij^{th} term in matrix $\mathbf{B}_p$, $\oslash$ denotes the pointwise division of two vectors. We can see that k_{HM} takes a mixture model form on its GSD.

2.4 Convolutional spectral kernels

Apart from a strict construction from the harmonizability definition, we can construct a non-stationary spectral kernel by convolving stationary spectral kernels, which gives a kernel form that is not only non-stationary, but also *nonparametric* in the sense that it takes functions as parameters.

Consider a feature map $K_{\mathbf{x}_i}(\mathbf{u})$ (see (2.12)) parametrized by a *linear span* of spectral mixture kernels:

$$K_{\mathbf{x}_i}(\mathbf{u}) = \sum_{q=1}^{Q} w_i^q \exp(-2\pi^2 S_i^q + 2i\pi\theta_i^q), \tag{2.26}$$

where $S_i^q = (\mathbf{u}-\mathbf{x}_i)^\top \mathbf{\Lambda}_i^q (\mathbf{u}-\mathbf{x}_i)$, $\theta_i^q = \langle \boldsymbol{\mu}_i^q, \mathbf{u}-\mathbf{x}_i\rangle$, and $\mathbf{\Lambda}_i^q$, $\boldsymbol{\mu}_i^q$ are input-dependent precision matrices and frequencies, and $\mathbf{\Sigma}_i = \mathbf{\Lambda}_i^{-1}$ are input-dependent input covariances. The convolution results in a total of Q^2 integrals:

$$\begin{aligned} k(\mathbf{x}_i, \mathbf{x}_j) &= \int_{\mathbb{R}^D} K_{\mathbf{x}_i}(\mathbf{u}) \overline{K}_{\mathbf{x}_j}(\mathbf{u}) \, \mathrm{d}\mathbf{u} \\ &= \sum_{1\le q,p\le Q} w_i^q \bar{w}_j^p \int \exp(-2\pi^2(S_i^q + S_j^p) + 2i\pi(\theta_i^q - \theta_j^p)) \, \mathrm{d}\mathbf{u}. \end{aligned} \tag{2.27}$$

The integral (2.27) can be solved by normalizing into a Gaussian density:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{1\le q,p\le Q} \frac{w_i^p \overline{w}_j^q}{(2\pi)^{D/2} |\mathbf{\Lambda}_i^q + \mathbf{\Lambda}_j^p|^{1/2}} \exp(-\pi^2 S_{i,j}^{qp} + 2i\pi\theta_{i,j}^{qp} - R_{i,j}^{qp}), \tag{2.28}$$

$$R_{i,j}^{qp} = (\boldsymbol{\mu}_i^q - \boldsymbol{\mu}_j^p)^\top \big((\mathbf{\Lambda}_i^q + \mathbf{\Lambda}_j^p)/2\big)^{-1} (\boldsymbol{\mu}_i^q - \boldsymbol{\mu}_j^p), \tag{2.29}$$

$$S_{i,j}^{qp} = (\mathbf{x}_i - \mathbf{x}_j)^\top \left((\mathbf{\Sigma}_i^q + \mathbf{\Sigma}_j^p)/2\right)^{-1} (\mathbf{x}_i - \mathbf{x}_j), \tag{2.30}$$

$$\theta_{i,j}^{qp} = \langle \mathbf{\Lambda}_i^q (\mathbf{\Lambda}_i^q + \mathbf{\Lambda}_j^p)^{-1} \boldsymbol{\mu}_j^p + \mathbf{\Lambda}_j^p (\mathbf{\Lambda}_i^q + \mathbf{\Lambda}_j^p)^{-1} \boldsymbol{\mu}_i^q, \mathbf{x}_i - \mathbf{x}_j \rangle. \tag{2.31}$$

We then employ normalization similar to that of Paciorek [20], and a positive definite matrix $\mathbf{B} \succeq \mathbf{0}_{Q\times Q}$:

$$k_{\text{conv}}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{1\le q,p\le Q} \frac{b_{qp} w_i^q \overline{w}_j^p |\mathbf{\Sigma}_i^q|^{\frac{1}{4}} |\mathbf{\Sigma}_j^p|^{\frac{1}{4}}}{|(\mathbf{\Sigma}_i^q + \mathbf{\Sigma}_j^p)/2|^{\frac{1}{2}}} \exp(-2\pi^2 S_{i,j}^{qp} + 2i\pi\theta_{i,j}^{qp} - R_{i,j}^{qp}). \tag{2.32}$$

k_{conv} (2.32) gives a *convolutional spectral kernel*.

2.4.1 Spectral interpretations

Unlike HMK, CSK does not have an intuitive spectral representation, but we can view the functions being convolved, $K_{\mathbf{x}_i}(\mathbf{u})$, as the "square root" of the local spectrum.

We can draw an equivalence between *characteristic functions* of probability distributions (or more generally, positive finite measures) and stationary kernels, for they are both the inverse Fourier transform of a finite measure. The input-dependent spectral measure is then characterized by the Fourier transform of $\mathcal{F}(K_{\mathbf{x}_i}) = \widehat{K}_{\mathbf{x}_i}$:

$$\widehat{K}_{\mathbf{x}_i}(\boldsymbol{\xi}) = \sum_{q=1}^{Q} w_i^q \mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu}_i^q, \boldsymbol{\Lambda}_i^q). \tag{2.33}$$

Because the Fourier transform of the convolution (2.27) is the product of the Fourier transform of the two functions being convolved. Therefore, we can see the convolution encodes a spectrum:

$$\widehat{K}_{\mathbf{x}_i,\mathbf{x}_j} = \widehat{K}_{\mathbf{x}_i}\widehat{K}_{\mathbf{x}_j}. \tag{2.34}$$

CSK can be interpreted as a "spectral mixture kernel" on a semi-metric space [29], where the distance between two points are defined by $d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{S_{i,j}^{qp}}$, which gives the exponential term of CSK a typical Gaussian kernel form $exp(d^2(\mathbf{x}_i, \mathbf{x}_j))$, and the covariance and frequencies are determined by the product of two local "square root of" spectra defined by (2.34).

2.5 Expressiveness of spectral kernels

In this section, we will discuss in detail the expressiveness of various spectral kernels. The discussion is motivated by an attempt to identify the exact expressive power of the spectral approach. We manage to demonstrate that certain types of spectral kernels are *dense* in the set of stationary, or harmonizable covariance functions, with the topology defined by pointwise approximation of functions.

2.5.1 Stationary spectral kernels

We will start with the discussion about existing stationary spectral kernels, namely the spectral mixture kernel [39]. We first demonstrate a generalized

version of the SM kernel, called the *generalized spectral kernel* in Samo and Roberts [28]:

$$k_{GS}(\boldsymbol{\tau}) = \sum_{q=1}^{Q} w_q^2 h(\boldsymbol{\tau} \circ \boldsymbol{\gamma}_q) \exp(2i\pi \boldsymbol{\omega}_q^\top \boldsymbol{\tau}), \tag{2.35}$$

where $h(\cdot)$ is a continuous, integrable stationary kernel. We propose the following theorem regarding (2.35):

Theorem 2. *Let h be a complex-valued positive definite, continuous and integrable function. Then the family of* generalized spectral kernels (2.35) *is dense in the family of stationary, complex-valued kernels with respect to pointwise convergence of functions. Here $\circ$ denotes the Hadamard product, $\boldsymbol{\omega}_q \in \mathbb{R}^D$, $\boldsymbol{\gamma}_q \in \mathbb{R}_+^D$, $Q \in \mathbb{N}_+$.*

Proof. We know from the uniform convergence of random Fourier features [23], that for an arbitrary stationary kernel $k_0(\mathbf{x}, \mathbf{x}') = k_0(\mathbf{x} - \mathbf{x}')$, for all compact subset $\mathcal{M} \in \mathbb{R}^D$, and for all $\epsilon > 0$, there exists a feature map $\zeta_\omega(\mathbf{x}) = \left(w_q e^{2\pi \boldsymbol{\omega}_q^\top \mathbf{x}}\right)_{q=1}^{Q}$, such that $|\zeta_\omega(\mathbf{x})\overline{\zeta_\omega(\mathbf{x}')} - k_0(\mathbf{x} - \mathbf{x}')| < \epsilon$. The uniform convergence of random Fourier features suggests the expressiveness of a generalized form of sparse spectrum kernel $k_{\text{SS}}(\mathbf{x} - \mathbf{x}') = \sum_{q=1}^{Q} w_q^2 e^{2\pi \boldsymbol{\omega}_q^\top (\mathbf{x} - \mathbf{x}')}$.

For an arbitrary continuous, integrable kernel h, consider the function $\tilde{k}(\boldsymbol{\tau}) = \dfrac{h(\boldsymbol{\tau} \circ \boldsymbol{\gamma})}{h(\mathbf{0})} k_{\text{SS}}(\boldsymbol{\tau}), \boldsymbol{\gamma} \succeq \mathbf{0}$. Because of the continuity of function h, $\tilde{k}$ uniformly approximates k_{SS} as $\boldsymbol{\gamma} \to \mathbf{0}^+$, and thus can be used to approximate any stationary covariance k_0.

$\tilde{k}(\boldsymbol{\tau})$ uniformly approximates any stationary kernel k_0 on arbitrary compact subset, $\mathcal{M}$ of $\mathbb{R}^D$. We can therefore construct a sequence of $\tilde{k}_n$ by setting $\epsilon_n = \dfrac{1}{n}$, $\mathcal{M}_n = \mathcal{B}(0, n) = \{v | \, \|v\| \leq n\}$, $n = 1, 2, 3, \cdots$. $\{\tilde{k}_n\}_{n=1}^{\infty}$ converges pointwise to k_0. k_{GS} takes a more general form, and thus has the same level of expressiveness as $\tilde{k}$. □

Theorem 2 speaks to the expressive power of the spectral mixture kernel: given an arbitrary stationary covariance k_0, we can find a parametrization of (2.35) that approximates k_0 with arbitrary precision. Samo and Roberts [28] present similar to Theorem 2, but their proposed family of "kernels" include invalid ones. However, later in Samo [27], a similarly strengthened result is presented. By presenting a strengthened result, we demonstrate that stationary covariances can be arbitrarily well approximated by a sequence of strictly valid SM kernels.

Kernel	Parameterization	Non-stationary	Non-monotonic	Nonparametric	Reference
SE	$Q = 1, \mathbf{\Lambda}^1 \equiv \text{diag}(1/\ell_{i,d}^2)_{d=1}^D, w_i^1 \equiv w_0$	✗	✗	✗	—
Gibbs	$Q = 1, \mathbf{\Lambda}_i^1 = \text{diag}(1/\ell_{i,d}^2)_{d=1}^D, w_i^1 \equiv 1$	✓	✗	✓	Gibbs [8]
PS	$Q = 1, \mathbf{\Lambda}_i^1 = \mathbf{\Sigma}_i^{-1}, w_i^1 \equiv 1$	✓	✗	✓	Paciorek and Schervish [19]
SM	$\mathbf{B} = \boldsymbol{I}, w_i^q \equiv w^q$	✗	✓	✗	Wilson and Adams [39]
GCSM‡	$\mathbf{B} = \mathbf{1}, w_i^q \equiv \sqrt{w^q}$	✗	✓	✗	Chen et al. [3]
HMK†	$w_i^q = \exp(2i\pi\langle \boldsymbol{\mu}^q, \mathbf{x}_i \rangle)$	✓	✓	✗	current work
CSK	—	✓	✓	✓	current work

Table 2.1: Subsets of CSK. The parametrization of HMK† corresponds to a variant form where h is stationary. The parametrization of the GCSM‡ corresponds to a subtype of GCSM with $\boldsymbol{\phi} = 0$, $\boldsymbol{\theta} = 0$. The full version is harmonizable and a subset of CSK, with a non-constant w_i^q.

In the proof of Theorem 2, we can see that the sparse spectrum (SS) kernel shares the same level of expressiveness with the SM kernel. However, sparse spectrum kernel encodes a finite dimensional feature map, which renders the kernel equivalent to a finite basis expansion. The SM kernel improves upon the SS kernel by adding uncertainty to the frequency components, which translates to an *integrable* kernel: instead of one frequency components propagating throughout the entirety of input space, the SM kernel gives a measure of regularization, with the added benefit of having infinite-dimensional features.

2.5.2 Non-stationary spectral kernels

In this section, we demonstrate the expressive power of *non-stationary spectral kernels*[2]. We demonstrate that the HMK's density in harmonizable covariances is similar to that of k_{GS} (2.35) in stationary covariances. Building on the expressiveness of HMK, we can determine the expressiveness of CSK by showing its ability to include several different cases of proposed expressive kernels as special cases.

We propose the following theorem with regards to the expressiveness of HMK:

Theorem 3. *Given a continuous, integrable kernel $h(\cdot, \cdot)$, the* harmonizable mixture kernel (2.22) *is dense in the family of harmonizable covariances with respect to pointwise convergence of functions.*

Proof. The proof follows a similar course as that of Theorem 3, with the observation that a generalized form of sparse spectrum kernel $\boldsymbol{\phi}(\mathbf{x})^\dagger \mathbf{B} \boldsymbol{\phi}(\mathbf{x}')$ is dense in the family of stationary covariances. □

[2]In this context, this phrase refers to the two non-stationary kernel family proposed in this thesis, and should not be confused with Remes et al. [25].

Theorem 3 depicts the expressive power of HMK, which suggests that HMK can almost approximate arbitrary bounded, continuous kernels.

While we cannot explicitly determine the expressiveness of CSK, we can show its ability to include various previously proposed kernels as special cases. The equivalence relationships are shown in table 2.1. CSK includes both non-spectral kernels (Gaussian, Gibbs and PS) as well as spectral kernels (SM, GCSM and HMK). It effectively spans harmonizable covariances for it includes variants of HMK as special cases, which is proved to approximate any harmonizable covariance. It is also not hard to see that CSK expands beyond harmonizable kernels, for the weight components w_q^i can be unbounded.

2.6 Interpreting spectral kernels

In this section, we discuss ways to interpret spectral kernels, namely using the Wigner distribution function (WDF) as a proxy for spectrogram.

2.6.1 Wigner distribution functions

A spectrogram shows the relation between input and frequency, however, this concept is less well-defined than the spectrum, which is usually obtained via a Fourier transform.

We suggest that Wigner distribution function [5] can be seen as a spectrogram. WDF is defined by the Wigner transform:

Definition 3. The *Wigner distribution function* (WDF) of a kernel $k(\cdot,\cdot) : \mathbb{R}^D \times \mathbb{R}^D \mapsto \mathbb{C}$ is defined as $W_k : \mathbb{R}^D \times \mathbb{R}^D \mapsto \mathbb{R}$:

$$W_k(\mathbf{x}, \boldsymbol{\omega}) = \int_{\mathbb{R}^D} k\left(\mathbf{x} + \frac{\boldsymbol{\tau}}{2}, \mathbf{x} - \frac{\boldsymbol{\tau}}{2}\right) e^{-2i\pi\boldsymbol{\omega}^\top\boldsymbol{\tau}}\, \mathrm{d}\boldsymbol{\tau}. \tag{2.36}$$

The Wigner transform first changes the kernel k into a function of the centroid of the input: $(\mathbf{x}+\mathbf{x}')/2$ and the lag $\mathbf{x}-\mathbf{x}'$, and then takes the Fourier transform of the lag. The Wigner distribution functions are fully equivalent to non-stationary kernels. Given the domain of WDF, we can view WDF as a 'spectrogram' demonstrating the relation between input and frequency. Converting an arbitrary kernel into its Wigner distribution sheds light into the frequency structure of the kernel.

The WDFs of locally stationary kernels adhere to the intuitive notion of local stationarity where frequencies remain constant at a local scale. Take locally stationary Gaussian kernel k_{LSG} (2.17) as an example:

$$W_{k_{\mathrm{LSG}}}(\mathbf{x}, \boldsymbol{\omega}) = \mathcal{N}(\boldsymbol{\omega}|\mathbf{0}, \boldsymbol{\Sigma}_2) e^{-2\pi^2\mathbf{x}^\top\boldsymbol{\Sigma}_1\mathbf{x}}. \tag{2.37}$$

There is, however, an important distinction between Wigner distribution and the conventional notion of a spectrogram. The spectrogram is usually viewed as a joint distribution between input and frequency, but Wigner distribution is characterized as a "quasiprobability distribution", which generalizes the probability density function by allowing for negative densities. While modeling on the domain of Wigner distributions seems promising in fully characterizing the non-stationary covariance functions, the effort is dampened by the inability to model negative densities.

The WDF of HMK (2.21) can be derived with a switch in subscript:

$$\begin{aligned} k_p(\mathbf{x}, \mathbf{x}') &= k_{\text{LSG}}(\mathbf{x} \circ \boldsymbol{\gamma}_p, \mathbf{x}' \circ \boldsymbol{\gamma}_p) \sum_{1 \le i,j \le Q_p} \beta_{pij} e^{2i\pi(\boldsymbol{\mu}_{pi}^\top \mathbf{x} - \boldsymbol{\mu}_{pj}^\top \mathbf{x}')} \\ &= k_{\text{LSG}}(\mathbf{x} \circ \boldsymbol{\gamma}_p, \mathbf{x}' \circ \boldsymbol{\gamma}_p)(Re(g(\widetilde{\mathbf{x}}, \boldsymbol{\tau})) + Im(g(\widetilde{\mathbf{x}}, \boldsymbol{\tau}))), \end{aligned} \tag{2.38}$$

$$Re(g(\widetilde{\mathbf{x}}, \boldsymbol{\tau})) = \sum_{1 \le i,j \le Q_p} \beta_{pij} \left(\cos 2\pi \left(\frac{\boldsymbol{\mu}_{pi} + \boldsymbol{\mu}_{pj}}{2} \right)^\top \boldsymbol{\tau} \right) \cos(2\pi(\boldsymbol{\mu}_{pi} - \boldsymbol{\mu}_{pj})^\top \widetilde{\mathbf{x}}), \tag{2.39}$$

$$Im(g(\widetilde{\mathbf{x}}, \boldsymbol{\tau})) = -Im(g(\widetilde{\mathbf{x}}, -\boldsymbol{\tau})). \tag{2.40}$$

The imaginary part has integral 0 with the Wigner transform, therefore we can derive the Wigner transform of HMK:

$$W_{k_{\text{HM}}}(\mathbf{x}, \boldsymbol{\omega}) = \sum_{p=1}^{P} W_{k_p}(\mathbf{x} - \mathbf{x}_p, \boldsymbol{\omega}), \tag{2.41}$$

$$W_{k_p}(\mathbf{x}, \boldsymbol{\omega}) = \frac{1}{\prod_{d=1}^{D} \gamma_{pd}} \sum_{1 \le i,j \le Q_p} W_{pij}(\mathbf{x}, \boldsymbol{\omega}), \tag{2.42}$$

$$W_{pij}(\mathbf{x}, \boldsymbol{\omega}) = W_{k_{\text{LSG}}}\left(\mathbf{x} \circ \boldsymbol{\gamma}_p, \left(\boldsymbol{\omega} - (\boldsymbol{\mu}_{pi} + \boldsymbol{\mu}_{pj})/2\right) \oslash \boldsymbol{\gamma}_p\right) \cos(2\pi(\boldsymbol{\mu}_{pi} - \boldsymbol{\mu}_{pj})^\top \mathbf{x}). \tag{2.43}$$

2.6.2 Spectral symmetry for real-valued kernels

In this chapter, we mainly discuss complex-valued kernel for the sake of generality. Real-valued kernels are a subset with certain 'spectral symmetry'. Such properties are easily detected when we see a real valued kernel as an average between a complex valued kernel and its conjugate: $k_r = \dfrac{k_c + \overline{k_c}}{2}$,

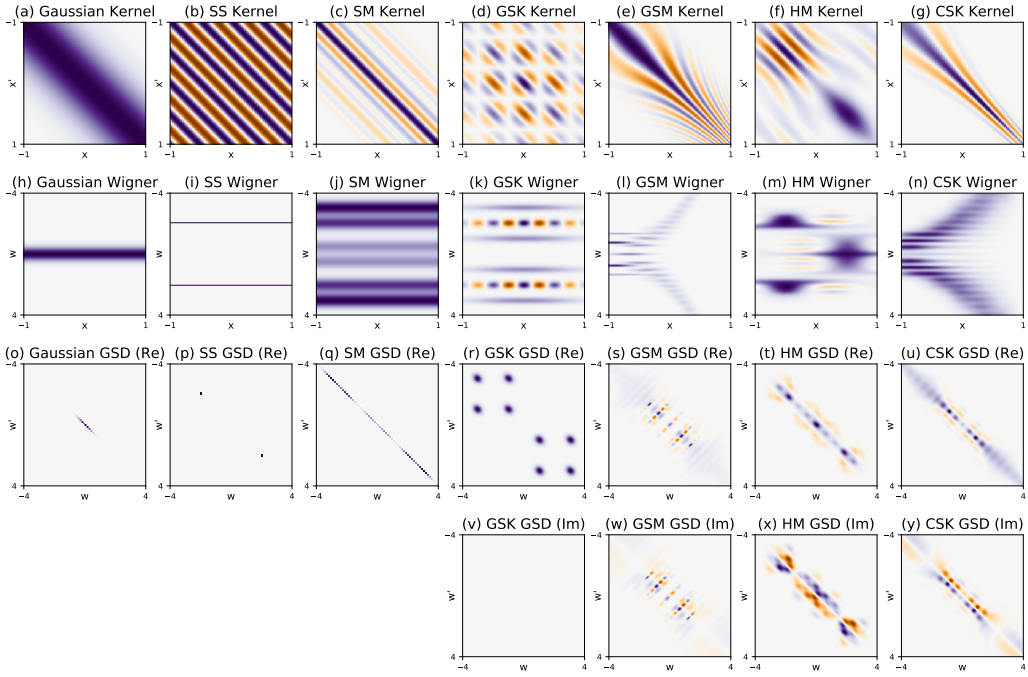

Figure 2.1: Overview of kernels and their spectral representations.

we can therefore derive the symmetry of its GSD and WDF:

$$W_{k_r}(\mathbf{x}, \boldsymbol{\xi}) = \frac{W_{k_c} + W_{\overline{k_c}}}{2} = \frac{W_{k_c}(\mathbf{x}, \boldsymbol{\xi}) + W_{k_c}(\mathbf{x}, -\boldsymbol{\xi})}{2}, \tag{2.44}$$

$$S_{k_r}(\boldsymbol{\omega}, \boldsymbol{\xi}) = \frac{S_{k_c} + S_{\overline{k_c}}}{2} = \frac{S_{k_c}(\boldsymbol{\omega}, \boldsymbol{\xi}) + S_{k_c}(-\boldsymbol{\omega}, -\boldsymbol{\xi})}{2}. \tag{2.45}$$

2.7 An overview of spectral kernels

We have extensively discussed the spectral properties of kernels. Figure 2.1 demonstrates an overall visualization of kernels and their spectral interpretations. As we can see from the first three columns of stationary kernels ((a), (b), (c)), stationary kernels are translation-invariant, and their Wigner distributions ((h), (i), (j)) are input-independent: $W(x, \omega) = W(x', \omega)$. The generalized spectral distribution ((o), (p), (q)) is concentrated on the diagonal[3].

As we can see from the non-stationary kernel matrices ((d), (e), (f), (g)), the kernel values are clearly input-dependent, with a clear rising trend for

[3]We have an intuitive visualization of the GSD for stationary kernels because stationary kernels do not allow a generalized spectral density

GSM and CSK. We can also observe the non-stationarity from the Wigner distributions ((k), (l), (m), (n)), where local spectra is not invariant with the input. We can see rising trend of the GSM and CSK, and we can observe the irregularity of Wigner distribution with figures (k) and (m), where the Wigner distributions involve negative densities. The frequencies of spectral kernels are clearly visible from the generalized spectral densities, where GSK and HMK have fixed frequency values, and GSM and CSK has a rising trend with respect to the input.

2.8 Kernel recovery experiments

We determined the expressiveness of spectral kernels in theory by proposing two new theorems in section 2.5. In this section, we demonstrate the expressiveness with empirical evidence.

2.8.1 Kernel recovery with HMK

From the theory in 2.5, we know that harmonizable mixture kernels are dense in harmonizable covariances, which include most bounded, continuous kernels. We use HMK to recover the kernel matrices of two non-stationary kernels, the covariance function of a time-inverted fractional Brownian motion (IFBM), and the generalized spectral kernel [25]:

$$k_{\text{IFBM}}(t,s) = \frac{1}{2}\left(\frac{1}{t^{2h}} + \frac{1}{s^{2h}} - \left|\frac{1}{t} - \frac{1}{s}\right|^{2h}\right), 0 < h < 1, \tag{2.46}$$

$$k_{\text{GSM}}(x,x') = w(x)w(x')k_{\text{Gibbs}}(x,x')\cos(2\pi(\mu(x)x - \mu(x')x')), \tag{2.47}$$

$$k_{\text{Gibbs}}(x,x') = \sqrt{\frac{2l(x)l(x')}{l^2(x)+l^2(x')}}\exp\left(-\frac{(x-x')^2}{l^2(x)+l^2(x')}\right). \tag{2.48}$$

We parametrize the Hurst index $h = 0.6$, and assign functional parameters $w(\cdot), l(\cdot), \mu(\cdot)$ with polynomials. We recover the kernel values in a compact subset of $\mathbb{R}$: $s,t \in (0.1,1]$, $x,x' \in [-1,1]$.

In 2.2, we can see that how HMK approximates IFBM and GSM kernels with low error.

2.8.2 CSK records unbiased frequency information

The *generalized spectral mixture kernel* (GSM) shares similar levels of expressiveness as convolutional spectral kernel. However, only CSK keeps an

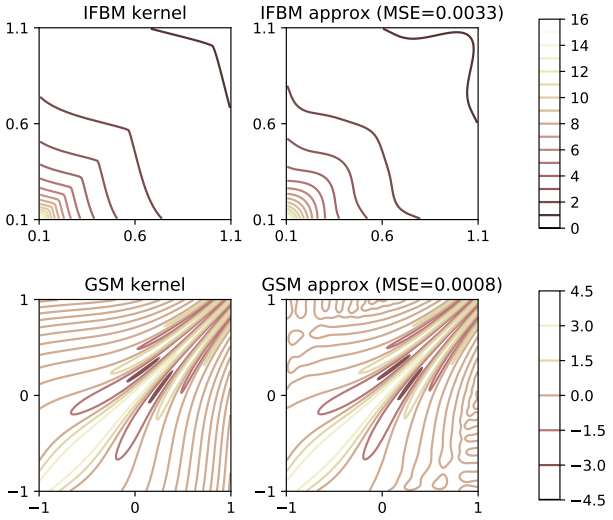

Figure 2.2: Kernel approximation of IFBM and GSM kernels.

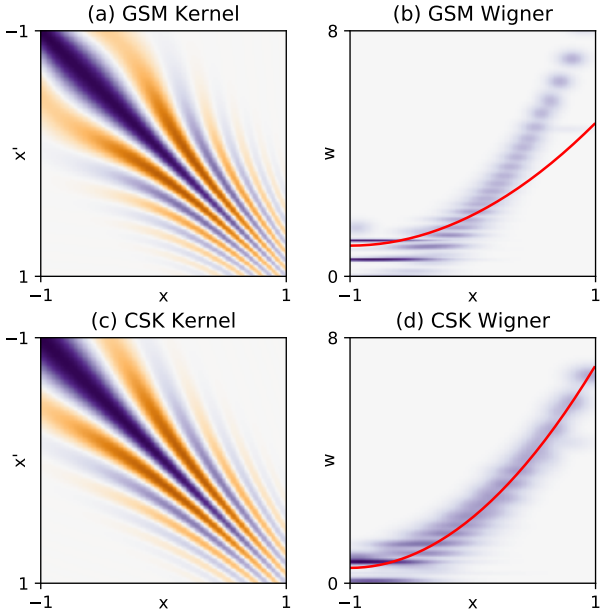

Figure 2.3: Kernel matrices and Wigner distribution functions of GSM and CSK, red lines denotes the values of the frequency function $\mu(\cdot)$.

unbiased record of frequency information with the frequency function parameter $\boldsymbol{\mu}$. In this section, we conduct kernel recovery experiments for GSM and CSK, which demonstrates that when two kernels share similar shape, the underlying frequency is accurately depicted with CSK as observed from the approximated WDF of the two kernels.

We use two different parametrizations of GSM and CSK, so that the two kernels share similar kernel values, and then we use HMK with $p = 20$ components to recover the kernel matrix. The result is demonstrated in Figure 2.3.

As we can see from Figure 2.3, the GSM kernel overestimates low frequencies and underestimates higher ones, while the frequency function values of CSK closely corresponds to the approximated WDF. The unbiasedness of CSK adds to the interpretability, where the function values of the frequency components can be directly interpreted with respect to the underlying periodicity of the function being inferred.

The distinction between GSM and CSK can also be easily seen from a first-order approximation of $\mu(\cdot)$. Assuming other functions remain constant, $\tilde{x} = (x + x')/2$, $\tau = x - x'$, then $\mu(x) \approx \mu(\tilde{x}) + \mu'(\tilde{x})\tau/2$. The cosine term for GSM is inaccurate in that the first-order approximation is $\cos(2\pi(\mu(x)x - \mu(x')x')) \approx \cos(2\pi\mu(\tilde{x})\tau + 2\pi\mu'(\tilde{x})(x^2 - x'^2))$, where the quadratic term is a bias. The same issue does not apply to CSK, where the first-order approximation of the cosine term is unbiased.

2.9 Summary

In this chapter, we discussed properties of harmonizable kernels, and propose two new classes of kernels, where the harmonizable mixture kernel is derived by applying a mixture model on its generalized spectral density, and the convolutional spectral kernel is derived from convolving two spectral mixture kernels as a feature map. Both kernels show high level of expressiveness given by theoretical and empirical evidence. The scope of discussion in this chapter is strictly contained within the kernel method framework, but we will discuss efficient inference using the proposed kernels in the next chapter.

Chapter 3

Inference

In this chapter, we study inference methods specifically designed for Gaussian process models with spectral kernels. We begin with a brief introduction to Gaussian processes and sparse variational inference in 3.1. We introduce *variational Fourier features*, an inter-domain sparse GP inference approach for *harmonizable mixture kernels* in 3.2. We derive *random Fourier features* for non-stationary kernels in 3.3, and we briefly introduce sparse inference for *convolutional spectral kernels* in 3.4.

3.1 Background

Gaussian processes (GP) are a popular nonparametric probabilistic framework noted for its nonlinearity, tractability and robustness to overfitting [24]. However, the power of GP models is hindered by the computation of the log-likelihood, which involves a matrix inversion with time complexity $O(N^3)$, where N is the number of data points. Recent studies have focused on the *scalability* of GP models, the most popular method being the sparse variational inference with pseudo-inputs [32], or *inducing points*. In this section, we introduce GPs and variational inference relevant to this thesis.

3.1.1 Gaussian processes

Gaussian processes are a generalization of the Gaussian distribution, which defines a collection of random variables, any finite number of which have a joint Gaussian distribution [24]. A GP is fully specified by the *mean function* and *covariance function*, or *kernel*. The mean function $m(\cdot)$ and covariance

function $k(\cdot,\cdot)$ of a real process $f(\mathbf{x})$ define the following expectations:

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})], \tag{3.1}$$
$$k(\mathbf{x},\mathbf{x}') = \mathbb{E}\big[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))\big]. \tag{3.2}$$

A GP prior is characterized as $f \sim \mathcal{GP}(m(\cdot), k(\cdot,\cdot))$. Without loss of generality, we study zero-mean GPs, i.e., $m(\cdot) \equiv 0$. A GP regression is a GP prior coupled with a Gaussian observation model with noise variance σ^2: $\mathbf{y}|\mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma^2\mathbf{I})$. We can subsequently derive the log-likelihood of a GPR model:

$$\log p(\mathbf{y}|\boldsymbol{\theta}) = -\frac{1}{2}\underbrace{\mathbf{y}^\top(\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2\mathbf{I})^{-1}\mathbf{y}}_{\text{data fit}} - \frac{1}{2}\underbrace{\log|\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2\mathbf{I}|}_{\text{model complexity}} - \frac{N}{2}\log 2\pi. \tag{3.3}$$

Here $\boldsymbol{\theta}$ represents the set of possible kernel hyperparamters. We can see from (3.3) that the log-likelihood naturally combines the data fit and model complexity, and we can optimize the hyperparameters $\boldsymbol{\theta}$ using likelihood maximization.

Because of the tractability of multivariate Gaussian distribution, we can easily derive the predictive distribution over the test set $\mathbf{X}_*$:

$$\begin{pmatrix}\mathbf{y}\\ \mathbf{f}_*\end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix}k(\mathbf{X},\mathbf{X}) + \sigma^2\mathbf{I} & k(\mathbf{X},\mathbf{X}_*)\\ k(\mathbf{X}_*,\mathbf{X}) & k(\mathbf{X}_*,\mathbf{X}_*)\end{pmatrix}\right), \tag{3.4}$$
$$\mathbf{f}_*|\mathbf{X},\mathbf{X}_*,\mathbf{y} \sim \mathcal{N}\left(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*\right), \tag{3.5}$$
$$\boldsymbol{\mu}_* = k(\mathbf{X}_*,\mathbf{X})\big(k(\mathbf{X},\mathbf{X}) + \sigma^2\mathbf{I}\big)^{-1}\mathbf{y}, \tag{3.6}$$
$$\boldsymbol{\Sigma}_* = k(\mathbf{X}_*,\mathbf{X}_*) - k(\mathbf{X}_*,\mathbf{X})\big(k(\mathbf{X},\mathbf{X}) + \sigma^2\mathbf{I}\big)^{-1}k(\mathbf{X},\mathbf{X}_*). \tag{3.7}$$

More generally, the posterior distribution is a GP with transformed mean and covariance functions:

$$\mathbf{f}(\cdot)|\mathbf{X},\mathbf{y} \sim \mathcal{GP}(\widehat{m}(\cdot), \widehat{k}(\cdot,\cdot)), \tag{3.8}$$
$$\widehat{m}(\mathbf{x}) = k(\mathbf{x},\mathbf{X})\big(k(\mathbf{X},\mathbf{X}) + \sigma^2\mathbf{I}\big)^{-1}\mathbf{y}, \tag{3.9}$$
$$\widehat{k}(\mathbf{x},\mathbf{x}') = k(\mathbf{x},\mathbf{x}') - k(\mathbf{x},\mathbf{X})\big(k(\mathbf{X},\mathbf{X}) + \sigma^2\mathbf{I}\big)^{-1}k(\mathbf{X},\mathbf{x}'). \tag{3.10}$$

3.1.2 Variational inference with inducing points

Variational inference (VI) is a Bayesian inference technique that transforms inference into an optimization problem [15, 37]. The main idea of VI is to approximate a possibly intractable posterior distribution, denoted by $p(\boldsymbol{\theta}|\boldsymbol{\alpha})$

where $\boldsymbol{\alpha}$ denotes possible hyperparameters, with a parametrized family of distributions, or *the approximating variational family* $\mathcal{Q} = \left\{q : q(\boldsymbol{\theta}; \eta), \eta \in \Omega_\eta\right\}$, minimizing the Kullback-Leibler (KL) divergence, which gives *the variational approximation* $q_{\boldsymbol{\alpha}}(\boldsymbol{\theta})$, the member in $\mathcal{Q}$ with the smallest KL-divergence:

$$q_{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \arg\min_{q \in \mathcal{Q}} \left\{KL(q(\boldsymbol{\theta}; \eta) \parallel p_{\boldsymbol{\alpha}}(\boldsymbol{\theta}))\right\}. \tag{3.11}$$

Minimization of the KL-divergence is equivalent to maximization of the evidence lower bound (ELBO), which is a function of the parameters of the approximated distributions,

$$\mathcal{L}(\eta) = \mathbb{E}_{q(\boldsymbol{\theta};\eta)} \left[\log p(\boldsymbol{\theta}, \boldsymbol{\alpha}) - \log q(\boldsymbol{\theta}; \eta)\right]. \tag{3.12}$$

In the case of Gaussian process regression, the posterior distribution $p(\mathbf{f}|\mathbf{X}, \mathbf{y})$ (3.8) involves inverting the matrix $k(\mathbf{X}, \mathbf{X}) + \sigma^2\mathbf{I}$, which has time complexity $O(N^3)$, where N is the number of training data points. Gaussian process models are computationally expensive due to the cubic time complexity of matrix inversion.

Various efforts have been applied to making GP models more scalable using variational inference, most notably, reducing the rank of the covariace matrix using sparse Gaussian processes [10, 22, 32, 34]. Sparse GP models approximate the posterior (3.8) by imposing an approximating variational family of a vector of pseudo inputs (or *inducing points*) $\mathbf{Z}$ and its functional values $\mathbf{u}$: $q(\mathbf{u}) = \mathcal{N}(\mathbf{m}, \mathbf{S})$, and minimize the distance between the true posterior $p(\mathbf{f}, \mathbf{u}|\mathbf{X}, \mathbf{y})$ and the approximated posterior $q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f}|\mathbf{u})q(\mathbf{u})$, which leads to an ELBO:

$$\mathcal{L}_1(\mathbf{Z}, \mathbf{m}, \mathbf{S}) = \mathbb{E}_{q(\mathbf{f},\mathbf{u})} \left[\log p(\mathbf{y}|\mathbf{f})p(\mathbf{u}) - \log q(\mathbf{u})\right]. \tag{3.13}$$

In GP regression, the lower bound (3.13) can be first maximized by analytically solving the variational parameters $(\mathbf{m}, \mathbf{S})$, which leads to the variational lower bound in Titsias [34]:

$$\mathcal{L}_2(\mathbf{z}) = \log \mathcal{N}(\mathbf{y}|0, k(\mathbf{X}, \mathbf{X}) - \mathbf{K}_{\mathbf{X}|\mathbf{Z}}) - \frac{1}{2\sigma^2}\mathrm{tr}(\mathbf{K}_{\mathbf{X}|\mathbf{Z}}), \tag{3.14}$$

where the matrix $\mathbf{K}_{\mathbf{X}|\mathbf{Z}}$ is the posterior covariance matrix $\mathbf{K}_{\mathbf{X}|\mathbf{Z}} = k(\mathbf{X}, \mathbf{X}) - k(\mathbf{X}, \mathbf{Z})k(\mathbf{Z}, \mathbf{Z})^{-1}k(\mathbf{Z}, \mathbf{X})$. The analytical solution of the variational parameters are given in Hensman et al. [10]:

$$\widehat{\mathbf{S}} = \frac{1}{\sigma^2}\mathbf{K}_{\mathbf{Z},\mathbf{Z}}^{-1}\mathbf{K}_{\mathbf{Z},\mathbf{X}}\mathbf{K}_{\mathbf{X},\mathbf{Z}}\mathbf{K}_{\mathbf{Z},\mathbf{Z}}^{-1} + \mathbf{K}_{\mathbf{Z},\mathbf{Z}}^{-1}, \tag{3.15}$$

$$\widehat{\mathbf{m}} = \frac{1}{\sigma^2}\widehat{\mathbf{S}}^{-1}\mathbf{K}_{\mathbf{Z},\mathbf{Z}}^{-1}\mathbf{K}_{\mathbf{Z},\mathbf{X}}\mathbf{y}. \tag{3.16}$$

The variational lower bound (3.14) only requires matrix inversion of size $m \times m$, where m is the number of inducing points. Considering $m \ll N$, computing the lower bound requires time complexity $O(Nm^2)$.

3.1.3 Variational inference with inducing features

The variational lower bound (3.14) requires computing covariances between the function values at data points and inducing points, i.e., $\mathbf{cov}(f(\mathbf{x}_i), f(\mathbf{z}_j))$, $\mathbf{cov}(f(\mathbf{z}_j), f(\mathbf{z}_k))$. The inter-domain Gaussian processes [17] generalizes the computation of cross-covariances by applying a linear transform, which projects the data into another domain where the pseudo dataset lies.

Consider a GP prior $f \sim \mathcal{GP}(0, k(\cdot,\cdot))$ and a linear transform $\mathscr{L}$, for instance, a convolution between $f(\mathbf{x})$ and a deterministic *feature extraction function* $g(\mathbf{x}, \mathbf{z})$:

$$\mathscr{L} f(\mathbf{z}) = \int f(\mathbf{x}) g(\mathbf{x}, \mathbf{z}) \mathrm{d}\mathbf{x}. \tag{3.17}$$

The function $h = \mathscr{L} f$ is a Gaussian process on the transformed domain, when the linear transform $\mathscr{L}$ is well defined with condition

$$\mathbf{cov}(h(\mathbf{z}), h(\mathbf{z}')) = \iint k(\mathbf{x}, \mathbf{x}') g(\mathbf{x}, \mathbf{z}) g(\mathbf{x}', \mathbf{z}') \mathrm{d}\mathbf{x} \mathrm{d}\mathbf{x}' < \infty. \tag{3.18}$$

The h function is a Gaussian process with zero mean and kernel defined in (3.18). The equation (3.18) also defines the cross-covariances between inducing "features" (as opposed to inducing points which lies on the same domain). The cross-covariances between input points and inducing features are also well-defined:

$$\mathbf{cov}(f(\mathbf{x}), h(\mathbf{z})) = \int k(\mathbf{x}, \mathbf{u}) g(\mathbf{u}, \mathbf{z}) \mathrm{d}\mathbf{z}. \tag{3.19}$$

Therefore, the variational lower bound can be computed with inter-domain Gaussian processes where the entries in $\mathbf{K}_{\mathbf{Z},\mathbf{X}}$ are replaced using (3.19), and the entries in $\mathbf{K}_{\mathbf{Z},\mathbf{Z}}$ are replaced using (3.18).

While the Fourier transform $g(\mathbf{x}, \mathbf{z}) = \exp(-2i\pi \mathbf{x}^\top \mathbf{z})$ is plausibly a good choice for a feature extraction function in that it provides inducing features on the feature domain, it generates invalid transformed GPs as the variances go to infinity for every frequency with the non-convergent integral in (3.18).

While inter-domain Gaussian processes [17] uses an ℓ_2 inner product between f and the feature extraction function g, other forms of inner product applies as well as long as the transform remain a well-defined Gaussian process. In Hensman et al. [12], the reproducing kernel Hilbert space (RKHS)

inner product i used to derive inter-domain harmonic inducing frequencies for Matérn type kernels, which gives a structured form of the cross-covariance matrix $K(\mathbf{Z}, \mathbf{Z})$, allowing for a linear complexity matrix inversion.

3.1.4 Sparse spectrum Gaussian processes

We introduced Bochner's theorem in 2.1. The equivalence between stationary kernels and finite measures can also be applied to obtaining scalable Gaussian process models.

A direct consequence of Bochner's theorem is a reformulation of stationary kernels as expectations:

$$\begin{aligned} k(\mathbf{x} - \mathbf{x}') &= k(\mathbf{0})\mathbb{E}_{\boldsymbol{\omega}\sim S_k}\left[\exp(2i\pi\boldsymbol{\omega}^\top(\mathbf{x} - \mathbf{x}'))\right] \\ &= k(\mathbf{0})\mathbb{E}_{\boldsymbol{\omega}\sim S_k}\left[\exp(2i\pi\boldsymbol{\omega}^\top\mathbf{x})\overline{\exp(2i\pi\boldsymbol{\omega}^\top\mathbf{x}')}\right]. \end{aligned} \tag{3.20}$$

The representation of stationary kernels as expectation of feature maps gives an approximation of kernels using random features [23]. Using dot product of random features for Gaussian process models, we obtain the sparse spectrum Gaussian processes [21], which approximates the original model using a generative model:

$$\boldsymbol{\omega}_j \sim S_k(\boldsymbol{\omega}), j = 1, \ldots, J, \tag{3.21}$$

$$\widehat{k}(\mathbf{x}, \mathbf{x}') = \frac{k(\mathbf{0})}{J}\sum_{j=1}^{J}\exp(2i\pi\boldsymbol{\omega}_j^\top(\mathbf{x} - \mathbf{x}')), \tag{3.22}$$

$$\mathbf{f}|\mathbf{X}, \boldsymbol{\Omega} \sim \mathcal{N}\big(\mathbf{0}, \widehat{k}(\mathbf{X}, \mathbf{X})\big), \boldsymbol{\Omega} = \big\{\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_J\big\}, \tag{3.23}$$

$$\mathbf{y}|\mathbf{f} \sim \mathcal{N}\big(\mathbf{f}, \sigma^2\mathbf{I}\big). \tag{3.24}$$

Sparse spectrum GP uses a finite-dimensional feature map, which makes the GP regression devolve into a Bayesian linear regression with trigonometric basis expansion [21], which has linear time complexity. Gal and Turner [6] improve sparse spectrum GP by placing variational distribution over the frequencies, which adds uncertainty measures in frequency inputs.

3.2 Variational Fourier features

In this section, we introduce *variational Fourier features* (VFF), an inter-domain inference approach designed for HMKs. We will first discuss the Fourier transform of GPs, and we will derive the VFF designed for the additive harmonizable mixture kernels (2.22).

3.2.1 Fourier transform of GPs

The Fourier transform is a linear transform decomposing a function of time in the frequency domain:

$$\mathcal{F}(f) = \hat{f}(\boldsymbol{\xi}) = \int_{\mathbb{R}^D} f(\mathbf{x}) e^{-2i\pi \boldsymbol{\xi}^\top \mathbf{x}} \, \mathrm{d}\mathbf{x}. \tag{3.25}$$

Previous research on inter-domain GPs focused on transforming the original GP with a Fourier transform: $\mathcal{L} = \mathcal{F}$, substituting inducing points with 'inducing frequencies'. However, such effort is dampened by the fact that sample paths from GPs with a stationary kernel are not square integrable *almost surely* (a.s.):

$$\mathbb{E}\big[\int |f(\mathbf{x})|^2 \, \mathrm{d}\mathbf{x} \big] = \int k(\mathbf{x}, \mathbf{x}) \, \mathrm{d}\mathbf{x} = \infty \tag{3.26}$$

Therefore, samples from a stationary GP do not have a Fourier transform. However, for integrable kernel k, the sample paths from $\mathcal{GP}(0, k)$ are square integrable almost surely, thus making the Fourier transform possible. The Fourier transform of a GP is a (complex-valued) GP:

$$\hat{f}(\boldsymbol{\omega}) \sim \mathcal{GP}(\hat{m}(\cdot), \hat{k}(\cdot, \cdot)), \tag{3.27}$$

$$\hat{m}(\boldsymbol{\omega}) = \mathbb{E}\left[\int f(\mathbf{x}) e^{-2\pi \boldsymbol{\omega}^\top \mathbf{x}} \, \mathrm{d}\mathbf{x} \right] = 0, \tag{3.28}$$

$$\begin{aligned} \hat{k}(\boldsymbol{\omega}, \boldsymbol{\xi}) &= \mathbb{E}\left[\iint f(\mathbf{x}) f(\mathbf{x}') e^{-2i\pi(\boldsymbol{\omega}^\top \mathbf{x} - \boldsymbol{\xi}^\top \mathbf{x}')} \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}' \right] \\ &= S_k(\boldsymbol{\omega}, \boldsymbol{\xi}), \end{aligned} \tag{3.29}$$

$$\begin{aligned} \mathbf{cov}(\hat{f}(\boldsymbol{\omega}), f(\mathbf{x})) &= \mathbb{E}\left[\int f(\mathbf{x}) f(\mathbf{t}) e^{-2i\pi \boldsymbol{\omega}^\top \mathbf{t}} \, \mathrm{d}\mathbf{t} \right] \\ &= \int k(\mathbf{t}, \mathbf{x}) e^{-2i\pi \boldsymbol{\omega}^\top \mathbf{t}} \, \mathrm{d}\mathbf{t}. \end{aligned} \tag{3.30}$$

Therefore, we can see that the transformed GP has a covariance function equivalent to the GSD of the original covariance function, and all integrals are proper because the kernel k is integrable.

The derivation is closed-form for HMK, the cross-covariances in (3.30) for

$k = k_p$ (2.21) is derived as follows:

$$\begin{aligned}\int k_p(\mathbf{t}, \mathbf{x}) e^{-2i\pi \boldsymbol{\xi}^\top \mathbf{t}} \, \mathrm{d}\mathbf{t} = &\sum_{1 \le i,j \le Q_p} \beta_{pij} \exp\left(-2\pi^2 \mathbf{x}^\top \left(\frac{\boldsymbol{\Sigma}_1}{4} + \boldsymbol{\Sigma}_2\right) \mathbf{x}\right) \\ &\times \exp\left(-2i\pi(\boldsymbol{\mu}_{pj}^\top \mathbf{x} + \boldsymbol{\xi}^\top \mathbf{x}_0)\right) \\ &\times \mathcal{N}\left((\boldsymbol{\xi} - \boldsymbol{\mu}_{pi}) \oslash \gamma_p \,\middle|\, 0, \frac{\boldsymbol{\Sigma}_1}{4} + \boldsymbol{\Sigma}_2\right), \qquad (3.31) \\ \mathbf{x}_0 = &(\boldsymbol{\Sigma}_1 + 4\boldsymbol{\Sigma}_2)^{-1}(4\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)\mathbf{x}. \qquad (3.32)\end{aligned}$$

3.2.2 Variational Fourier features of harmonizable mixture kernel

HMK belongs to the kernel family discussed in 3.2.1, but we can further utilize the additive structure of an HMK, $k_{HM} = \sum_{p=1}^{P} k_p(\mathbf{x} - \mathbf{x}_p, \mathbf{x}' - \mathbf{x}_p)$. A GP with kernel k_{HM} can be decomposed into P independent GPs:

$$f(\mathbf{x}) = \sum_{p=1}^{P} f_p(\mathbf{x} - \mathbf{x}_p), \qquad (3.33)$$

$$f_p(\mathbf{x}) \sim \mathcal{GP}(0, k_p(\mathbf{x}, \mathbf{x}')). \qquad (3.34)$$

Given this formulation, we can derive *variational Fourier features* with inducing frequencies conditioned on one f_p. For the p^{th} component, we have m_p inducing frequencies $(\boldsymbol{\omega}_{p1}, \ldots, \boldsymbol{\omega}_{pm_p})$ and m_p inducing values $(u_{p1}, \cdots, u_{pm_p})$. We can compute inter-domain covariances in a similar fashion:

$$\begin{aligned}\mathbf{K}_{fu}(\boldsymbol{\omega}_{qj}, \mathbf{x}) &\triangleq \mathbf{cov}(f(x), u_{qj}) \qquad (3.35) \\ &= \sum_{p=1}^{P} \mathbf{cov}(f_p(\mathbf{x} - \mathbf{x}_p), u_{qj}) \\ &= \mathbf{cov}(f_q(\mathbf{x} - \mathbf{x}_q), \hat{f}_q(\boldsymbol{\omega}_{qj})).\end{aligned}$$

Similarly, we compute the entries of the matrix K_{uu}

$$\mathbf{K}_{uu}(\boldsymbol{\omega}_{pi}, \boldsymbol{\omega}_{qj}) \triangleq \mathbf{cov}(u_{pi}, u_{qj}) = \begin{cases} S_p(\boldsymbol{\omega}_{pi}, \boldsymbol{\omega}_{qj}), p = q, \\ 0, p \neq q. \end{cases} \qquad (3.36)$$

The matrix $\mathbf{K}_{uu}$ has a block diagonal structure, which allows for faster matrix inversion. The variational Fourier features are then completed by plugging in the entries in $\mathbf{K}_{fu}$ (3.35) and $\mathbf{K}_{uu}$ (3.36) into the evidence lower bound (3.13).

3.3 Random Fourier features

In this section, we derive the random Fourier features for HMK, which can be generalized to other harmonizable covariances. We then use the generative model to derive another variational inference paradigm for HMK. Here we consider (2.21), where the entries in matrix $\mathbf{B}_p$ are nonnegative, which makes the GSD S_p a proper probability distribution, making it possible to sample tuples of frequency $(\boldsymbol{\omega}, \boldsymbol{\xi})$ from S_p[1]. We can rewrite (2.22).

$$k_{\mathrm{HM}}(\mathbf{x}, \mathbf{x}') = \sum_{p=1}^{P} \sigma_p^2 k_p(\mathbf{x} - \mathbf{x}_p, \mathbf{x}' - \mathbf{x}_p), \tag{3.37}$$

$$k_p(\mathbf{x}, \mathbf{x}') = k_{\mathrm{LSG}}(\mathbf{x} \circ \boldsymbol{\gamma}_p, \mathbf{x}' \circ \boldsymbol{\gamma}_p)\phi_p(\mathbf{x})^\top \mathbf{B}_p \phi_p(\mathbf{x}'), \tag{3.38}$$

where $\mathbf{1}^\top \mathbf{B}_p \mathbf{1} = 1$. Given the generalized Fourier transform defined in (2.14), we can find the following unbiased estimate of k_p:

$$k_p(\mathbf{x}, \mathbf{x}') = \sigma_p^2 \mathbb{E}_{(\boldsymbol{\omega}, \boldsymbol{\xi}) \sim S_p} \left[e^{2i\pi(\boldsymbol{\omega}^\top \mathbf{x} - \boldsymbol{\xi}^\top \mathbf{x}')} \right]. \tag{3.39}$$

However, this estimate does not translate to a feature map, which renders the estimate $e^{2i\pi(\boldsymbol{\omega}^\top \mathbf{x} - \boldsymbol{\xi}^\top \mathbf{x}')}$ not positive definite with probability 1.

We can use N draws from $S_p(\boldsymbol{\omega}, \boldsymbol{\xi})$. Denote $\boldsymbol{\Omega} = \{\boldsymbol{\omega}_n^1, \boldsymbol{\omega}_n^2 | n = 1, 2, \ldots, N\}$:

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= \sigma_p^2 \lim_{N \to \infty} \frac{1}{4N^2} \mathbb{E}_{(\boldsymbol{\omega}_n^1, \boldsymbol{\omega}_n^2) \sim S_p} \left[\sum_{\boldsymbol{\omega}, \boldsymbol{\xi} \in \boldsymbol{\Omega}} e^{2i\pi(\boldsymbol{\omega}^\top \mathbf{x} - \boldsymbol{\xi}^\top \mathbf{x}')} \right] \\ &= \lim_{N \to \infty} \mathbb{E}_{(\boldsymbol{\omega}_n^1, \boldsymbol{\omega}_n^2) \sim S_p} \left[\phi_N(\mathbf{x}) \overline{\phi_N(\mathbf{x}')} \right], \end{aligned} \tag{3.40}$$

$$\phi_N(\mathbf{x}) = \frac{\sigma_p}{2N} \sum_{\boldsymbol{\omega} \in \boldsymbol{\Omega}} e^{2i\pi \boldsymbol{\omega}^\top \mathbf{x}}. \tag{3.41}$$

Using multiple draws from GSD, we can construct a consistent, albeit biased estimate of the kernel. For real-valued kernels, the complex feature map ϕ_N (3.41) can be rewritten as a two-dimensional feature map:

$$\varphi_N(\mathbf{x}) = \frac{\sigma_p}{2N} \begin{pmatrix} \sum_{\boldsymbol{\omega} \in \boldsymbol{\Omega}} \cos(2\pi \boldsymbol{\omega}^\top \mathbf{x}) \\ \sum_{\boldsymbol{\omega} \in \boldsymbol{\Omega}} \sin(2\pi \boldsymbol{\omega}^\top \mathbf{x}) \end{pmatrix}. \tag{3.42}$$

Introducing an auxiliary variable in the same fashion as Rahimi and Recht [23], we can rewrite (3.42):

$$\varphi_{N,b}(\mathbf{x}) = \frac{\sigma_p}{\sqrt{2}N} \sum_{\boldsymbol{\omega} \in \boldsymbol{\Omega}} \cos(2\pi \boldsymbol{\omega}^\top \mathbf{x} + b), b \sim \mathrm{Unif}[0, 2\pi]. \tag{3.43}$$

[1] It is imperative to note the fact that S_p is not necessarily a valid probability distribution.

A random Fourier feature is constructed by stacking (3.43):

$$\boldsymbol{\psi}_p(\mathbf{x}) = \frac{1}{\sqrt{M}} \begin{pmatrix} \varphi_{N_1,b_1}(\mathbf{x}) \\ \cdots \\ \varphi_{N_m,b_m}(\mathbf{x}) \\ \cdots \\ \varphi_{N_M,b_M}(\mathbf{x}) \end{pmatrix}. \tag{3.44}$$

An immediate generalization of random Fourier features is to formulate as a generative model, and represent uncertainty in the frequency inputs [6]. However, this approach is not feasible due to the computation of ELBO involving an intractable KL-divergence term between two Gaussian mixtures.

3.4 Sparse inference for CSK

In this section, we study sparse inference for convolutional spectral kernel for bivariate or univariate data. We impose Gaussian process priors over the "component functions" $w^q(\cdot)$, $\mu_d^q(\cdot)$ and $\boldsymbol{\Lambda}^q(\cdot) = \begin{pmatrix} \lambda_1^2 & \rho\lambda_1\lambda_2 \\ \rho\lambda_1\lambda_2 & \lambda_2^2 \end{pmatrix}$:

$$w^q(\cdot) \sim \mathcal{GP}(c_w, k_{\text{SE}}(\cdot,\cdot)), \tag{3.45}$$
$$\text{logit}(\mu_d^q(\cdot)) \sim \mathcal{GP}(c_\mu, k_{\text{SE}}(\cdot,\cdot)), d = 1, 2, \tag{3.46}$$
$$\lambda_d(\cdot) \sim \mathcal{GP}(c_{\lambda_d}, k_{\text{SE}}(\cdot,\cdot)), d = 1, 2, \tag{3.47}$$
$$\text{logit}(\rho(\cdot)) \sim \mathcal{GP}(c_\rho, k_{\text{SE}}(\cdot,\cdot)). \tag{3.48}$$

The hierarchical model consists of the latent Gaussian process function $f(\mathbf{x})$ and the three component functions $w(\mathbf{x})$, $\boldsymbol{\mu}(\mathbf{x})$, $\Lambda(\mathbf{x})$ as kernel functions. We construct a sparse variational inference approximation of the latent function [11], and estimate a MAP solution for the component functions. We parameterize the variational approximation with shared inducing locations $\mathbf{Z}$ with inducing point distribution $q(\mathbf{u}_f) = \mathcal{N}(\mathbf{u}_f|\mathbf{m}, \mathbf{S})$. The component function's values are determined by inducing points $\mathbf{U} = (\mathbf{u}_w, \mathbf{u}_\mu, \mathbf{u}_\Lambda)$, and with corresponding kernel lengthscales $\boldsymbol{\ell} = (\ell_w, \ell_\mu, \ell_\Lambda)$ and variances $\boldsymbol{\sigma} = (\sigma_w, \sigma_\mu, \sigma_\Lambda)$.

Following Hensman et al. [11], we infer the latent function posterior approximation by minimising the ELBO,

$$\arg\max_{\mathbf{U},\mathbf{Z},\boldsymbol{\ell},\boldsymbol{\sigma}} \Big\{ \log p(\mathbf{y}) \geq \mathbb{E}_{q(\mathbf{f})} \log p(\mathbf{y}|\mathbf{f})] - \text{KL}[q(\mathbf{u}_f)||p(\mathbf{u}_f)] \Big\} \tag{3.49}$$
$$q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{A}\mathbf{m}, \mathbf{K}_{\mathbf{XX}} - \mathbf{A}(\mathbf{K}_{\mathbf{ZZ}} - \mathbf{S})\mathbf{A}^\top), \tag{3.50}$$

where the kernel function depends on the three component functions. The model parameters are $(\mathbf{m}, \mathbf{S}, \mathbf{Z}; \mathbf{u}_w, \mathbf{u}_\mu, \mathbf{u}_\Lambda; \boldsymbol{\ell}, \boldsymbol{\sigma})$. We model the bivariate precision matrix with a bivariate Gaussian with a correlation ρ. where

$$q(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{m}, \mathbf{S}) \tag{3.51}$$

$$q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{A}\mathbf{m}, \mathbf{K}_{XX} - \mathbf{A}(\mathbf{K}_{ZZ} - \Sigma)\mathbf{A}^\top) \tag{3.52}$$

3.5 Summary

In this chapter, we applied the two classes of spectral kernels as covariance functions for Gaussian process models. GPs with HMK sample square integrable sample paths, which allows for the construction of *variational Fourier features*, an inter-domain sparse inference paradigm with inducing frequencies. The introduction of phase shifts $\mathbf{x}_p$ turns the generalized spectral density into a proper density function, which allows for *random Fourier features* for non-stationary kernels, a consistent biased estimate of kernel values. Finally, the inference for CSK is done by shared inducing points location of the kernel component functions, and the function being inferred. We will test the efficacy of the inference in the next chapter.

Chapter 4

Experiments

In this chapter, we demonstrate the power of spectral kernels in the context of classification and regression tasks. We show that spectral kernels manage to extract interpretable patterns within the framework of Gaussian process models. For harmonizable mixture kernels, we use a simplied version of the harmonizable kernel where the two matrices of the locally stationary k_{LSG} are diagonals: $\mathbf{\Sigma}_1 = \text{diag}(\sigma_d^2)$, $\mathbf{\Sigma}_2 = \lambda^2 I$. For convolutional spectral kernels, we also use a simplified version where the subscripts in (2.28) only contain items where $q = p$. This is equivalent to using a slightly different convolution with concatenation of feature vectors:

$$K^q_{\mathbf{x}_i}(\mathbf{u}) = w_i^q \exp(-2\pi^2 S_i^q + 2i\pi\theta_i^q), q = 1, \ldots, Q, \tag{4.1}$$

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{q=1}^{Q} \int_{\mathbb{R}^D} K^q_{\mathbf{x}_i}(\mathbf{u}) K^q_{\mathbf{x}_j}(\mathbf{u}). \tag{4.2}$$

The variational framework using inducing frequencies is based on Titsias [34] for GP regression, Hensman et al. [10] for GP classification, Salimbeni et al. [26] for natural gradient optimization of variational parameters.

4.1 GP classification

In this section, we test the efficacy of HMK in GP classification with the banana dataset, with similar settings as in Hensman et al. [11]. We show the effectiveness of variational Fourier fetures in GP classification with HMK. We use an HMK with $P = 4$ components to classify the banana dataset, and compare SVGP with inducing points (IP) [11] and SVGP with variational Fourier features (VFF). The model parameters are learned by alternating optimization rounds of natural gradients for the variational parameters, and

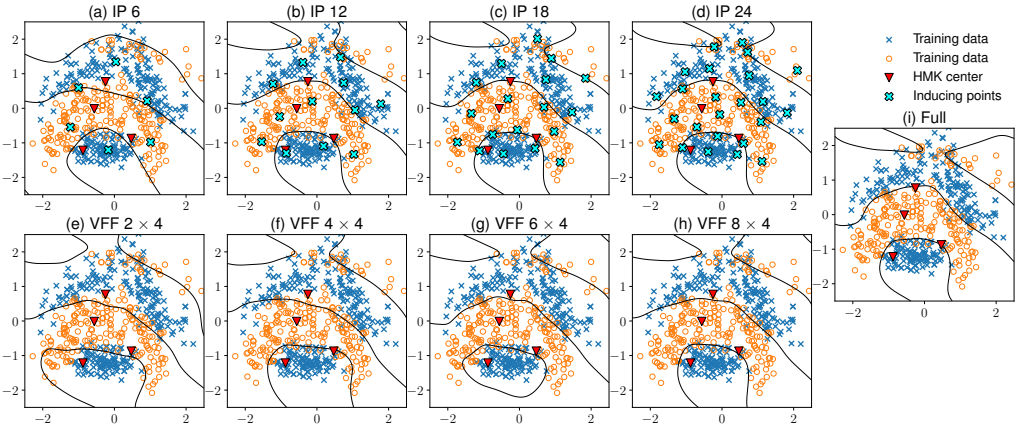

Figure 4.1: Sparse GP classification with the banana dataset. The model is learned by an HMK with $P = 4$ components, and thus 2 inducing frequencies for each component constitute a total of 2×4 inducing frequencies.

Adam optimizer for the other parameters [26]. Figure 4.1 shows the decision boundaries of the two methods over the number of inducing points. For both variants, we experiment with model complexities from 6 to 24 inducing points in IP, and from 2 to 8 inducing frequencies for each component of HMK in the VFF. The centers of HMK (red triangles) spread to support the data distribution. The IP method is slightly more complex compared to VFF at the same parameter counts in terms of nonzero entries in the variational parameters.

The VFF method recovers roughly the correct decision boundary even with a small number of inducing frequencies, while converging faster to the decision boundaries as the number of inducing frequencies increases.

4.2 GP regression

4.2.1 Harmonizable mixture kernel

In this section, we demonstrate the effectiveness of HMK in interpolation for the non-stationary solar irradiance dataset. We run sparse GP regression with squared exponential, spectral mixture and harmonizable mixture kernels, and show the predicted mean, and 95% confidence intervals for each model (See Figure 4.2). We use sparse GP regression proposed in [34] with 50 inducing points marked at the x axis. The SE kernel cannot estimate the periodic pattern and overestimates the signal smoothness. The SM kernel fits

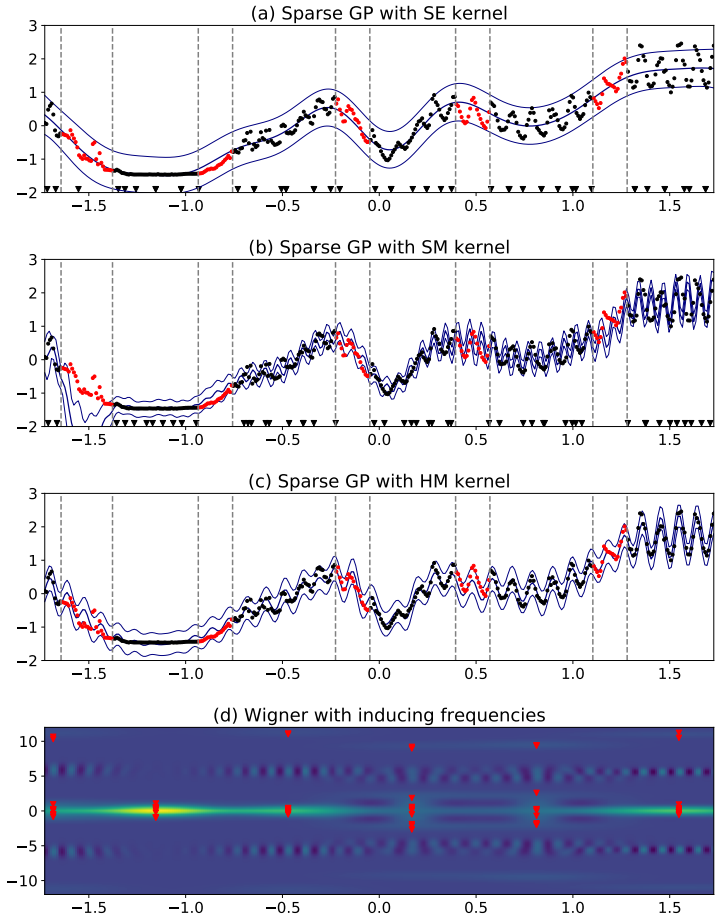

Figure 4.2: Sparse GP regression with solar irradiance dataset.

the training data well, but misidentifies frequencies on the first and fourth interval of the test set.

For sparse GP with HMK, we use the same framework where the variational lower bound is adjusted for VFF. The model extrapolates better for the added flexibility of nonstationarity, and the inducing frequencies aggregate near the learned frequencies. Both first and last test intervals are well fitted. The Wigner distribution with inducing frequencies of the optimised HM kernel is shown in Figure 4.2d.

4.2.2 Convolutional spectral kernel

We replicate the experimental setting for convolutional spectral kernels. We observe good predictive performance on the test data, and sensible trends

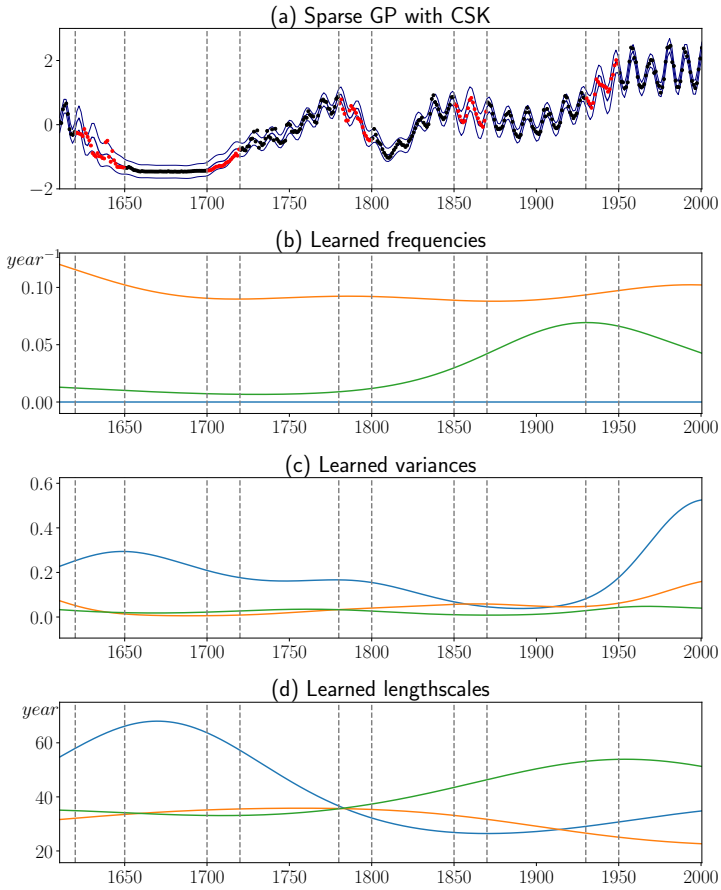

Figure 4.3: CSK with solar irradiance. Both the training points (black) and test points (red) are fitted accurately (a). The frequency, variance and lengthscale functions have three components (b-d).

for component functions.

GP regression with CSK exhibits better performance compared with HMK. While HMK "interpolates" local patterns by summing over the different components, while CSK gets rid of this rigid framework by effectively calculating adaptive frequencies for each pair of data points using a weighted average of the frequency function values.

Chapter 5

Discussion

In this chapter, we discuss the possible pitfalls of spectral kernels, the most important of which being the possibility of overfitting.

While an automated search of expressive kernels for Gaussian process models has proven successful on a variety of occasions, it is paramount to notice the effect of overfitting, which has been a largely overlooked issue in current research. In this chapter, we discuss how stationary spectral kernels would overfit any functions without proper regularization.

5.1 Overfitting of sparse spectrum kernels

The sparse spectrum kernel [21] can be seen as a special case of spectral mixture kernel [39] when the covariance of frequency $\boldsymbol{\Sigma}_q = \mathbf{0}$,

$$k_{\mathrm{SS}}(\mathbf{x}, \mathbf{x}') = \sum_{q=1}^{Q} \alpha_q^2 \exp(2i\pi \boldsymbol{\omega}_q^\top (\mathbf{x} - \mathbf{x}')) \tag{5.1}$$

$$= \mathbf{w} \circ \boldsymbol{\phi}(\mathbf{x}) \cdot \overline{\mathbf{w} \circ \boldsymbol{\phi}(\mathbf{x}')}, \tag{5.2}$$

where $\phi(\mathbf{x})_q = \exp(2i\pi \boldsymbol{\omega}_q^\top \mathbf{x})$. Using the dot product representation, it is easy to see a GP regression $f \sim \mathcal{GP}(0, k_{\mathrm{SS}})$ is equivalent to the following Bayesian linear regression with a trigonometric basis expansion,

$$y = \boldsymbol{\beta}^\top \boldsymbol{\phi}(\mathbf{x}) + \epsilon, \tag{5.3}$$

$$\boldsymbol{\beta} \sim \mathcal{N}(0, \mathrm{diag}(\mathbf{w}^2)). \tag{5.4}$$

The flexibility of sparse spectrum kernels lies within the learnable frequencies $\boldsymbol{\omega}_q$ and prior variances on coefficients β_q.

The trigonometric basis expansion is indeed too flexible if we view $\phi(\mathbf{x})_q$ as a set of linearly independent functions in the separable Hilbert space

$L^2(\mathcal{M})$, where $\mathcal{M}$ is a compact subset of $\mathbb{R}^D$. We can find an orthonormal Fourier basis of $L^2(\mathcal{M})$, which implies that

$$\forall f \in L^2(\mathcal{M}), f(\mathbf{x}) = \sum_{q=1}^{\infty} \beta_q \exp(2i\pi \boldsymbol{\omega}_q^\top \mathbf{x}), \tag{5.5}$$

where the convergence is in the L^2 norm. Given a large enough $Q \in \mathbb{N}_+$, the Bayesian linear regression with basis expansion $\phi(\mathbf{x})_q = \exp(2i\pi \boldsymbol{\omega}_q^\top \mathbf{x})$ will approximate any $f \in L^2(\mathcal{M})$ with good enough precision.

The perspective of a trigonometric basis expansion sheds light on overfitting and optimization issues for sparse spectrum kernels. Sparse spectrum kernels manage to asymptotically span the entirety of square integrable functions defined on $\mathcal{M}$, which tends to include overfitting solutions to a regression. Sparse spectrum kernel can also fit arbitrary square integrable functions given a fixed set of frequencies values $\{\boldsymbol{\omega}_q\}_{q=1}^Q$, which negates the effect of interpretable kernel learning: the frequencies values are supposed to be the peaks of the spectrum of a stationary process.

5.2 Complexity of spectral mixture kernels

Spectral mixture kernels [39] attempts to remedy the caveats of sparse spectrum kernel. Instead of a degenerate kernel with a finite rank Q, the spectral mixture kernel multiplies each spectral component $\exp(2i\pi \boldsymbol{\omega}_q^\top (\mathbf{x}-\mathbf{x}'))$ with a Gaussian kernel (or other non-degenerate stationary kernel [28]), which translates into "uncertainty" of frequency components on the spectrum. Wilson [40] demonstrates that the multiplication of a non-degenerate kernel makes SM kernel more robust to overfitting compared with SS kernel. However, we think that such modification might be insufficient in that SM kernel demonstrates the counterintuitive fact that "simple models overfit".

When we maximize the log-likelihood of a Gaussian process:

$$\log p(\mathbf{y}|\boldsymbol{\theta}) = -\frac{1}{2}\underbrace{\mathbf{y}^\top (\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2 \mathbf{I})^{-1}\mathbf{y}}_{\text{data fit}} - \frac{1}{2}\underbrace{\log |\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2 \mathbf{I}|}_{\text{model complexity}} - \frac{N}{2}\log 2\pi, \tag{5.6}$$

we see the log-determinant of the noisy kernel matrix $\log |\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2 \mathbf{I}|$ as a measure of the model complexity, and therefore, the log-likelihood formula is a compromise between model fit and complexity. However, when the Gaussian process is equipped with a SM kernel:

$$k(\mathbf{x}, \mathbf{x}') = \sum_{q=1}^{Q} \alpha_q^2 \exp(-2\pi^2 (\mathbf{x} - \mathbf{x}')^\top \boldsymbol{\Sigma}_q (\mathbf{x} - \mathbf{x}') + 2i\pi \boldsymbol{\omega}_q^\top (\mathbf{x} - \mathbf{x}')). \tag{5.7}$$

When we decrease $|\mathbf{\Sigma}_q|$, the kernel matrix becomes less diagonally dominant with the diagonal values unchanged, which decreases $\log|\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2\mathbf{I}|$. Therefore, the maximization of log-likelihood gives a SM kernel whose value does not decay within long range, behaving similar to a sparse spectrum kernel.

The non-degeneracy of SM kernel might not provide the intended regularizing effect. One way of regularization is to introduce a process over all possible SM kernels [14].

Given that harmonizable covariances include stationary ones as special cases, the same argument can be made about HMK. We need to proceed with caution with non-stationary Gaussian process modeling given the above argument about overfitting.

Chapter 6

Conclusions

In this thesis, we discuss the theoretical and practical aspects of spectral kernels used in Gaussian processes. The main contributions of this thesis include:

- We introduced *harmonizability*, a concept previously only used in probability theory, into machine learning. Harmonizable kernels help provide new insight into the non-stationary modeling of stochastic processes.
- We propose *harmonizable mixture kernel* (HMK), a new kernel family that is non-stationary, non-monotonic and with closed-form tractable interpretations. We determine theoretically and empirically the expressive power of HMK.
- We propose *convolutional spectral kernel* (CSK), a new kernel family taking functions as kernel hyperparameters. We demonstrate that CSK has superior expressiveness as it contains multiple previously proposed expressive kernels as special cases.
- We propose *variational Fourier features* (VFF), an inter-domain sparse variational inference scheme designed for HMK. We show that HMK has the desirable property of producing sample paths that are most likely square integrable, which makes the VFF capable of recovering the *exact* kernel form.
- We propose *random Fourier features* (RFF), an algorithm that generates a biased, consistent estimate of harmonizable kernels.

While there is much to be done with respect to the proper regularization for spectral kernels, given its tendency to overfit the data, this thesis lays

the groundwork for non-stationary spectral modeling for Gaussian process models.

Chapter 7

Bibliography

[1] Cedric Archambeau and Francis Bach. Multiple Gaussian process models. *arXiv preprint arXiv:1110.5238*, 2011.

[2] Salomon Bochner. *Lectures on Fourier Integrals: With an Author's Supplement on Monotonic Functions, Stieltjes Integrals and Harmonic Analysis; Translated from the Original German by Morris Tenenbaum and Harry Pollard*. Princeton University Press, 1959.

[3] Kai Chen, Perry Groot, Jinsong Chen, and Elena Marchiori. Spectral mixture kernels with time and phase delay dependencies. *arXiv preprint arXiv:1808.00560*, 2018.

[4] Nicolas Durrande, James Hensman, Magnus Rattray, and Neil D. Lawrence. Detecting periodicities with Gaussian processes. *PeerJ Computer Science*, 2:e50, 2016.

[5] Patrick Flandrin. *Time-frequency/time-scale analysis*, volume 10. Academic press, 1998.

[6] Yarin Gal and Richard Turner. Improving the Gaussian process sparse spectrum approximation by representing uncertainty in frequency inputs. In *International Conference on Machine Learning*, pages 655–664, 2015.

[7] Marc G. Genton. Classes of kernels for machine learning: a statistics perspective. *Journal of Machine Learning Research*, 2(Dec):299–312, 2001.

[8] Mark N. Gibbs. *Bayesian Gaussian processes for regression and classification*. PhD thesis, Citeseer, 1998.

[9] Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12(Jul):2211–2268, 2011.

[10] James Hensman, Nicolo Fusi, and Neil D. Lawrence. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.

[11] James Hensman, Alexander G. de G. Matthews, and Zoubin Ghahramani. Scalable variational Gaussian process classification. *AISTATS*, 2015.

[12] James Hensman, Nicolas Durrande, and Arno Solin. Variational Fourier features for Gaussian processes. *The Journal of Machine Learning Research*, 18(1):5537–5588, 2017.

[13] D. Higdon, J. Swall, and J. Kern. Non-stationary spatial modeling. *Bayesian statistics*, 6(1):761–768, 1999.

[14] Phillip A. Jang, Andrew Loeb, Matthew Davidow, and Andrew G. Wilson. Scalable Lévy process priors for spectral kernel learning. In *Advances in Neural Information Processing Systems*, pages 3940–3949, 2017.

[15] Michael Irwin Jordan. *Learning in graphical models*, volume 89. Springer Science & Business Media, 1998.

[16] Y. Kakihara. A note on harmonizable and v-bounded processes. *Journal of Multivariate Analysis*, 16:140–156, 1985.

[17] Miguel Lázaro-Gredilla and Anibal Figueiras-Vidal. Inter-domain Gaussian processes for sparse inference using inducing features. In *Advances in Neural Information Processing Systems*, pages 1087–1095, 2009.

[18] Michel Loève. Probability theory II (graduate texts in mathematics), 1994.

[19] Christopher J. Paciorek and Mark J. Schervish. Nonstationary covariance functions for Gaussian process regression. In *Advances in Neural Information Processing Systems*, pages 273–280, 2004.

[20] Christopher Joseph Paciorek. *Nonstationary Gaussian processes for regression and spatial modelling*. PhD thesis, Citeseer, 2003.

[21] Joaquin Quiñonero Candela, Carl Edward Rasmussen, Aníbal R. Figueiras-Vidal, et al. Sparse spectrum Gaussian process regression. *Journal of Machine Learning Research*, 11(Jun):1865–1881, 2010.

[22] Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6(Dec):1939–1959, 2005.

[23] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2008.

[24] C.E. Rasmussen and K.I. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

[25] Sami Remes, Markus Heinonen, and Samuel Kaski. Non-stationary spectral kernels. In *Advances in Neural Information Processing Systems*, pages 4642–4651, 2017.

[26] Hugh Salimbeni, Stefanos Eleftheriadis, and James Hensman. Natural gradients in practice: Non-conjugate variational inference in Gaussian process models. *arXiv preprint arXiv:1803.09151*, 2018.

[27] Yves-Laurent Kom Samo. *Advances in kernel methods: towards general-purpose and scalable models*. PhD thesis, University of Oxford, 2017.

[28] Yves-Laurent Kom Samo and Stephen Roberts. Generalized spectral kernels. *arXiv preprint arXiv:1506.02236*, 2015.

[29] Isaac J Schoenberg. Metric spaces and completely monotone functions. *Annals of Mathematics*, pages 811–841, 1938.

[30] R. Silverman. Locally stationary random processes. *IRE Transactions on Information Theory*, 3(3):182–187, 1957.

[31] A.J. Smola and B. Schölkopf. *Learning with kernels*, volume 4. Citeseer, 1998.

[32] Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems*, pages 1257–1264, 2006.

[33] Shengyang Sun, Guodong Zhang, Chaoqi Wang, Wenyuan Zeng, Jiaman Li, and Roger Grosse. Differentiable compositional kernel learning for Gaussian processes. *arXiv preprint arXiv:1806.04326*, 2018.

[34] Michalis Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009.

[35] Felipe Tobar, Thang D. Bui, and Richard E. Turner. Learning stationary time series using Gaussian processes with nonparametric kernels. In *Advances in Neural Information Processing Systems*, pages 3501–3509, 2015.

[36] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[37] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[38] Christopher K.I. Williams. Computing with infinite networks. In *Advances in Neural Information Processing Systems*, pages 295–301, 1997.

[39] Andrew Wilson and Ryan Adams. Gaussian rocess kernels for pattern discovery and extrapolation. In *International Conference on Machine Learning*, pages 1067–1075, 2013.

[40] Andrew Gordon Wilson. *Covariance kernels for fast automatic pattern discovery and extrapolation with Gaussian processes*. PhD thesis, University of Cambridge, 2014.

[41] A. M. Yaglom. *Correlation theory of stationary and related random functions: Volume I: Basic results*. Springer Series in Statistics. Springer, 1987.